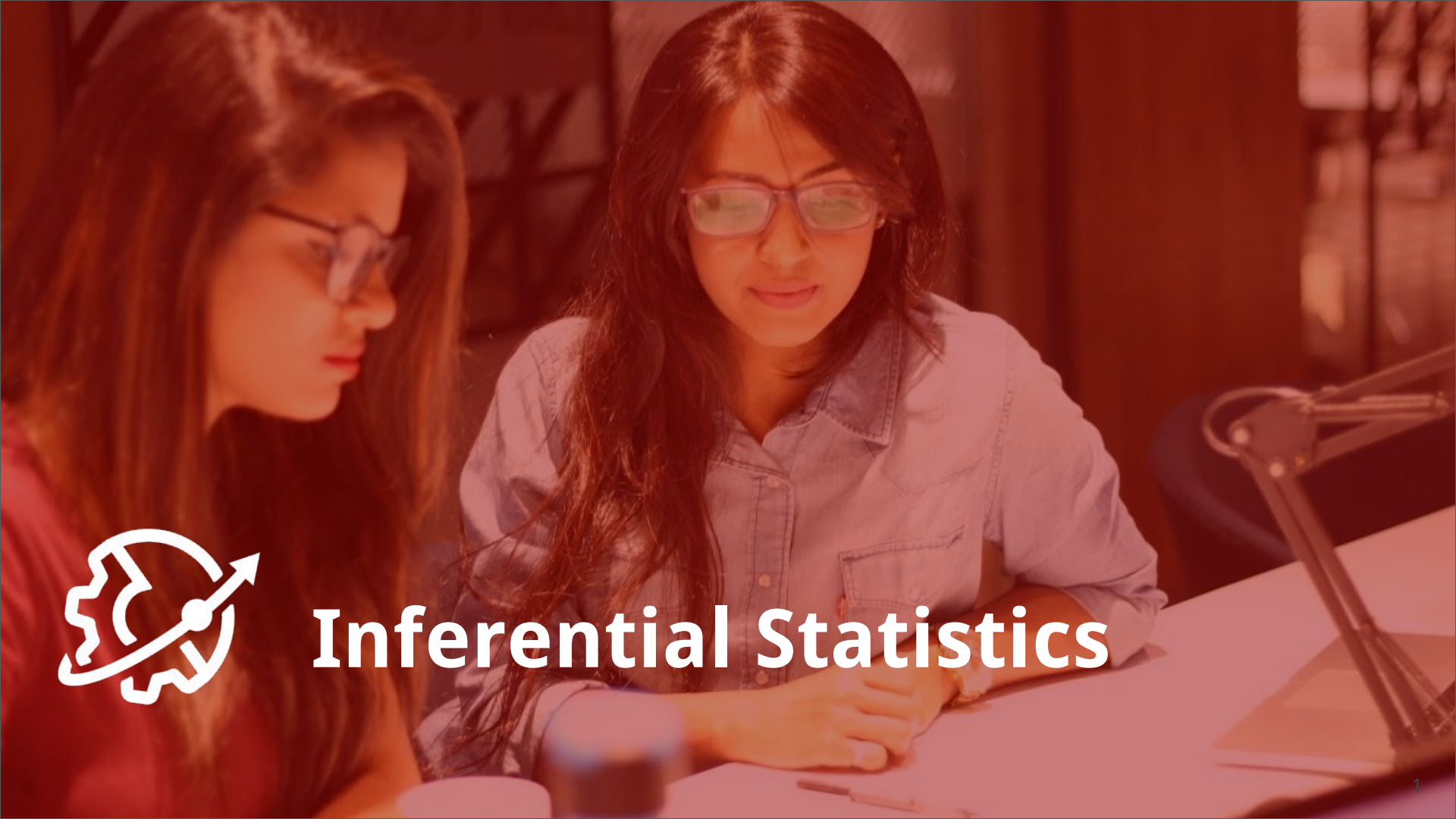


Inferential Statistics

Agenda

- Quick Review
- z-score
- Introduction to Statistical Inference
- Point Estimates and Confidence Intervals
- Hypothesis Testing

Quick Review

What is Inferential Statistics

We use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study.

Inferential statistics makes inferences about populations using data drawn from the population. Instead of using the entire population to gather the data, the statistician will collect a sample or samples from the millions of residents and make inferences about the entire population using the sample.

Inferential Statistics: Examples

Determine if the light-bulbs coming off the assembly line or faulty or not.

Determine if the lifetime of the light-bulbs coming off the assembly line changed (significantly) after a change in the manufacturing procedure.

Determine if the public thinks whether the prime-minister is doing a good job or not.

Descriptive vs Inferential Statistics

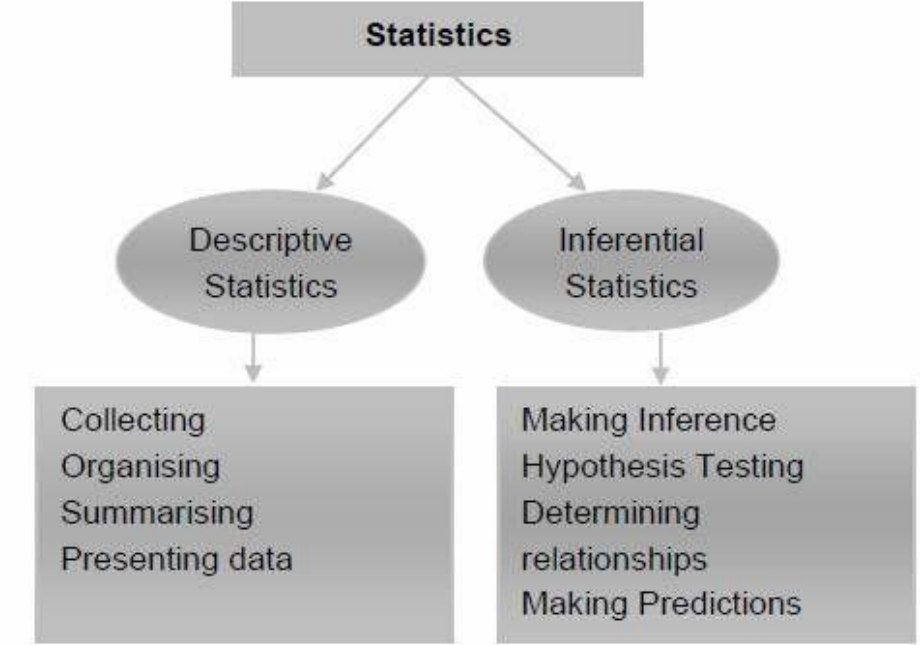
Descriptive statistics uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.

Inferential statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.

Descriptive vs Inferential Statistics

- Descriptive statistics is used only to describe the sample or summarise information about the sample.

- Inferential statistics use a random sample of data taken from a population to make inferences about the population.



Descriptive and Inferential Statistics

Suppose we are studying the heights of people of a population. After collecting their heights, we provide the tallest, shortest and average height of the population. This method of describing the population is called as **Descriptive Statistics.**

We then categorise the heights as ‘Tall’, ‘Short’, ‘Medium’ and take sample from the population to study/infer more about the sample so as to generalise about the population which is called as **Inferential Statistics.**

Why Inferential Statistics

Now that we know how to do feature engineering, we can generate hundreds of them. Depending on the computational power and time available to us and depending on the specific model we are using, it may often become crucial to select the right set of variables (**Feature Selection**)

- which have the maximum amount of predictive power and
- do not expose us to overfitting.

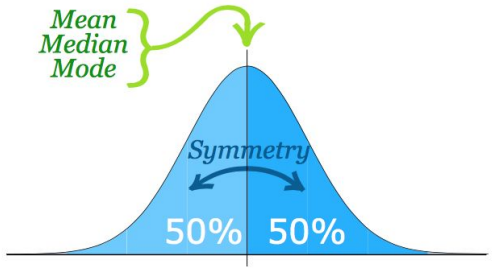
Inferential Statistics has many tools which are necessary in assessing the discriminative (predictive) power of the variables. So we shall take a small tour through Inferential Statistics so that we learn how to use these tools.

z-score

Normal Distribution

The normal or Gaussian distribution is a continuous probability distribution characterized by a symmetric bell-shaped curve. A normal distribution is defined by its center (mean) and spread (standard deviation).

The bulk of the observations generated from a normal distribution lie near the mean, which lies at the exact center of the distribution: as a rule of thumb, about 68% of the data lies within 1 standard deviation of the mean, 95% lies within 2 standard deviations and 99.7% lies within 3 standard deviations.

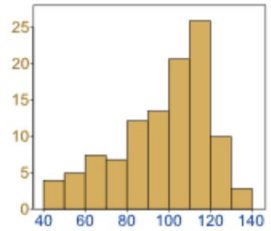


Properties of Normal Distribution:

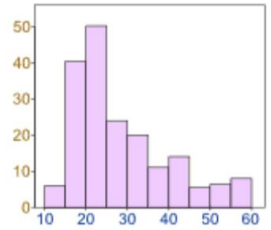
- Mean = median = mode.

Normal Distribution

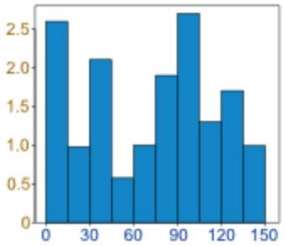
Data can be spread in different ways:



It can be concentrated more towards left.



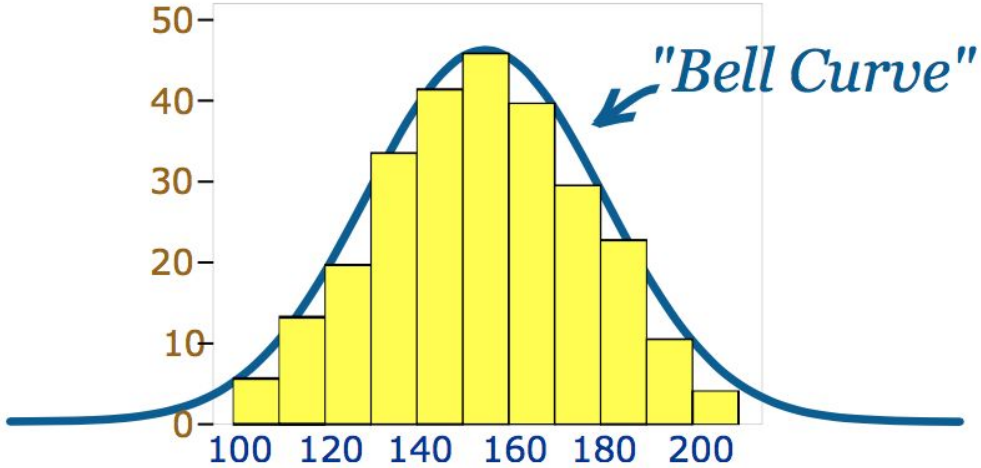
It can be concentrated more towards right.



It can be uneven.

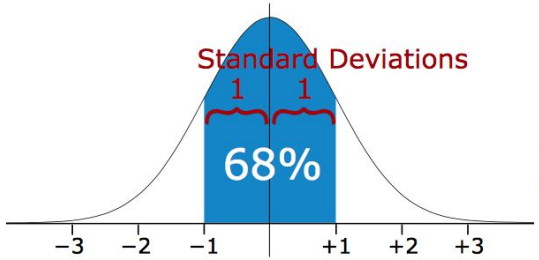
Normal Distribution

But there are instances where data tends to be around a central value with no bias on either sides and gets close to **bell shaped curve** like on right which is **Normal Distribution.**

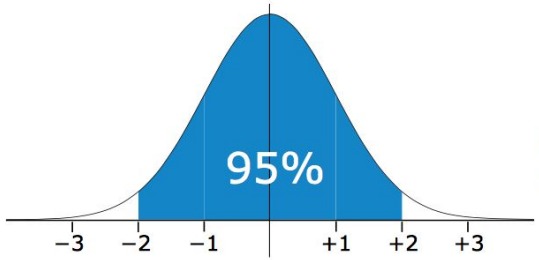


Normal Distribution

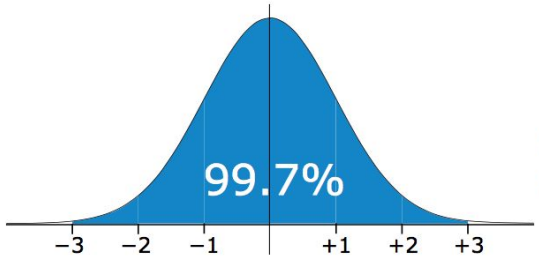
- An empirical rule of 68-95-99.7 is described on the right.



68% of values are within
1 standard deviation of the mean



95% of values are within
2 standard deviations of the mean



99.7% of values are within
3 standard deviations of the mean

Normal Distribution

- It has zero skew and kurtosis.
- $X \sim N(\mu, \sigma^2)$

Example: 95% of students at a college weigh between 55 kgs and 85 kgs. Assuming the data is normally distributed we need to calculate the mean and standard deviation.

Soln: Mean is halfway between 55 and 85 kgs = (55 + 85)/2 = 70 kgs.

95% is 2 standard deviations away from mean on each side, hence a total of 4 standard deviations.
Unit standard deviation = (85 - 55)/4 = 7.5 kgs

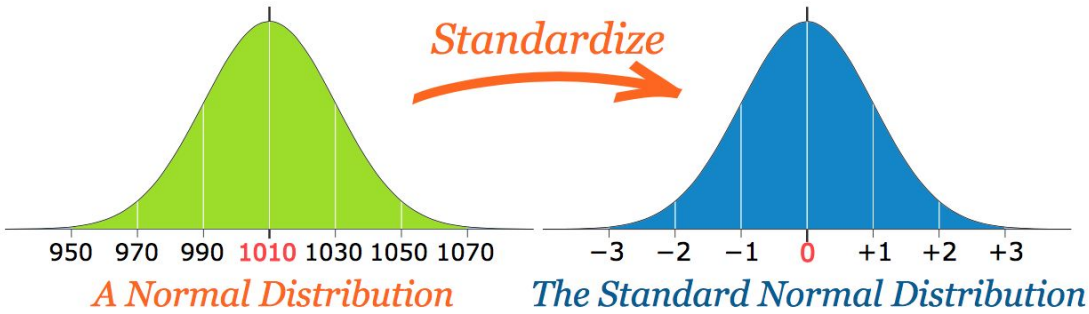
Knowing standard deviation is always good as we can say that any value is:
- likely to be within a standard deviation.
- very likely to be within 2 standard deviations.
- almost certainly within 3 standard deviations.

Z-Score/Standard Score

The number of standard deviations away from the mean is called the z-score or standard score.

Calculating z-score:
- Initially subtract the mean.
- Then divide by the standard deviation.

We can convert any normal distribution to standard normal distribution.



Normal Distribution

Continuing from the example, we meet a guy from the same college who weighs 92.5 kgs. We need to calculate his z-score.

Soln: We calculate how far is 92.5 from the mean

=> 92.5 - 70 = 22.5 kgs

How many standard deviations is 22.5 kgs ?

The standard deviation is 7.5 kgs so,

22.5/7.5 = 3 standard deviations.

Hence the guy is 3 standard deviations away from the mean.

Introduction to Statistical Inference

What is inference?

Statistical inference is the process of analyzing sample data to gain insight into the population from which the data was collected and to investigate differences between data samples.

Population & Sample: Example

- A census is a collection of complete data of entire population under study.

Example: 10 year census.

- A survey uses representative group of the given population to determine its characteristics.

Example: Opinion polls, quality control checks in manufacturing.

CENSUS 2011: RELIGIOUS PROFILE

India's population data based on religion, which was part of Census 2011, was released by the government on Tuesday

	Population in 2011 (cr)	Proportion of population in %	Decadal change in proportion in % pts
Hindu	96.63	79.8	-0.7
Muslim	17.22	14.2	+0.8
Christian	2.78	2.3	No change
Sikh	2.08	1.7	-0.2
Buddhist	0.84	0.7	-0.1
Jain	0.45	0.4	No change
Others	0.79	0.7	-
Religion not stated	0.29	0.2	-

Population vs Sample

It is not always convenient or possible to examine every member of an entire population and hence a subset of people or events is collected called as sample to infer about the entire population. It is the set of values that we use for estimation.

- To represent the population well, a sample should be randomly collected and adequately large.

- If the sample is random and large enough, you can use the information collected from the sample to make inferences about the population.

Example: It is not practical to count the bruises/dents on all apples picked at an orchard. It is possible, however, to count the bruises on a set of apples taken from that population. This subset of the population is called a sample.

Parameter & Statistic

- A parameter is a descriptive measure of the population.

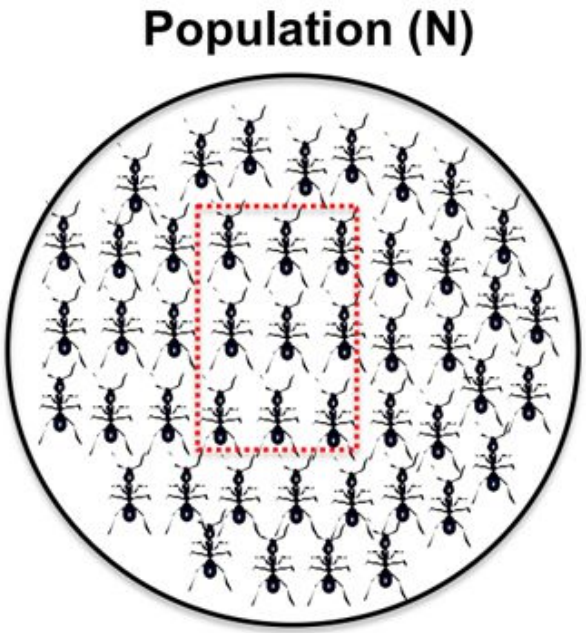
Example: Population mean, Population variance etc.

- A statistic is a descriptive measure of the sample.

Example: Sample mean, Sample variance etc.

Greek – Population Parameter	**Roman – Sample Statistic**
Mean – μ	Mean – $\bar{x}$
Variance – σ^2	Variance – s^2
Standard Deviation - σ	Standard Deviation - s

Population & Sample



Inference: Example

In data analysis, we are often interested in the characteristics of some large population, but collecting data on the entire population may be infeasible.

For example, leading up to U.S. presidential elections it could be very useful to know the political leanings of every single eligible voter, but surveying every voter is not feasible.

Instead, we could poll some subset of the population, such as a thousand registered voters, and use that data to make inferences about the population as a whole.

Techniques in Statistical Inference

1. **Confidence Intervals:** for estimating values of population parameters

2. **Hypothesis Testing:** for deciding whether the population supports a specific idea/model/hypothesis

Point Estimates and Confidence Intervals

Point Estimates

Point estimates are estimates of population parameters based on sample data.

For instance, if we wanted to know the average age of registered voters in the U.S., we could take a survey of registered voters and then use the average age of the respondents as a point estimate of the average age of the population as a whole.

More on Point Estimates

The sample mean is usually not exactly the same as the population mean. This difference can be caused by many factors including poor survey design, biased sampling methods and the randomness inherent to drawing a sample from a population.

Confidence Interval

A point estimate can give you a rough idea of a population parameter like the mean, but estimates are prone to error and taking multiple samples to get improved estimates may not be feasible.

A confidence interval is a range of values above and below a point estimate that captures the true population parameter at some predetermined confidence level.

Confidence Interval

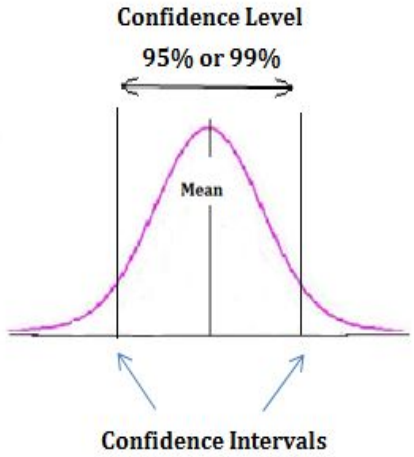
For example, if you want to have a 95% chance of capturing the true population parameter with a point estimate and a corresponding confidence interval, you'd set your confidence level to 95%. Higher confidence levels result in a wider confidence intervals.

Confidence Interval and Confidence Level

A **confidence interval** is a statistical way of saying that “I’m pretty sure that the true value of a number I am approximating is within this range. How sure am I? I am *s*% sure.” Where *s* is usually 95 or 99 and the range consists of numeric values.

When sampling from a population to estimate a mean a confidence interval is a range of values within which we are *s*% confident the true mean is included. *s* is some stated percentage, called a ***confidence level.***

If *s* = 95 then one can say, "In 95 out of 100 samples my estimated mean will fall within this stated range. Therefore, the true mean has a 95% chance of falling within this range. Conversely, there is a 5% chance that the true mean is not within this interval."



Calculating a Confidence Interval

- Calculate a confidence interval by taking a point estimate and then adding and subtracting a margin of error to create a range.

- Margin of error is based on your desired confidence level, the spread of the data and the size of your sample.

- The way you calculate the margin of error depends on whether you know the standard deviation of the population or not.

Calculating a Confidence Interval

If you know the standard deviation of the population, the CI is equal to:

$$Estimate \pm marginof error$$

$$Estimate \pm (z - value) * (SDof estimate)$$

Usually written as: (Lower confidence limit, Upper confidence limit)

$$(Est. - z * \times (SDof est.), Est. + z * \times (SDof est.))$$

Calculating a Confidence Interval

For Example: if σ is known to be 5, and we sample $\bar{x}$ to be 18, then the 95% confidence interval is:

$$18 +- 1.96 \times (5) = (8.2, 27.8)$$

Here, we have considered 95% of a Normal distribution between -1.96 and 1.96

Confidence Interval: Example

A survey was taken of German companies that do business with firms in India. One of the survey questions was: Approximately how many years has your company been trading with firms in India? A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation for this question is 7.7 years. Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in India for the population of German companies trading with firms in India.

Soln: n = 44, xbar = 10.455, σ = 7.7

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ or Margin of error} = z * \frac{\sigma}{\sqrt{n}}$$

$\therefore$ Confidence Interval for the Population Mean is

Sample Mean $\pm$ Margin of Error

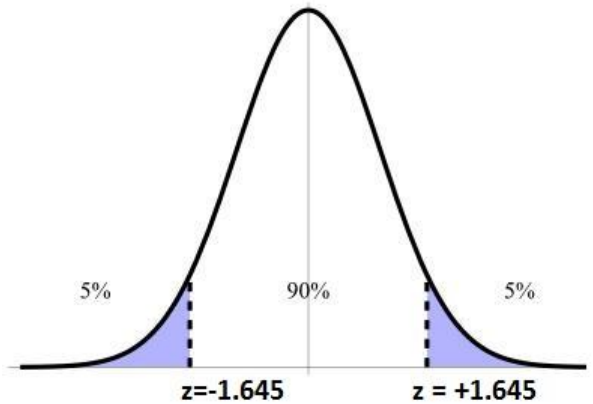
Confidence Interval: Example

z_a and z_b be the lower and upper confidence limits respectively.

$P(z_a < Z < z_b) = 0.9$

$P(Z < z_a) = 0.05$

$P(Z > z_b) = 0.05$



Confidence Interval: Example

Margin of error at 90% confidence level = 1.645 * (7.7/√44) = 1.91

$$\bar{X} - 1.91 < \mu < \bar{X} + 1.91$$

This is confidence interval for population mean: **Sample mean ± Margin of Error.**

Since the sample mean is 10.455 years, we get confidence interval for 90% as
8.545 < μ < 12.365

The analyst is 90% confident that if a census of all German companies trading with firms in India were taken at the time of the survey, the actual population mean number of trading years of such firms would be between 8.545 and 12.365 years.

Knowledge Check

Q.1) The standard score of value 25 is 3. If the variance is 16 then the mean value is ?

Q.2) The standard score for N(10, 9), value 7 is ?

Q.3) For a huge data set how will you infer whether it follows normal distribution without plotting ?

Knowledge Check

Q.1) The standard score of value 25 is 3. If the variance is 16 then the mean value is ?

Ans.) $3 = (25 - \mu)/4 \Rightarrow \mu = 13$.

Q.2) The standard score for N(10, 9), value 7 is ?

Q.3) For a huge data set how will you infer whether it follows normal distribution without plotting ?

Knowledge Check

Q.1) The standard score of value 25 is 3. If the variance is 16 then the mean value is ?

Ans.) $3 = (25 - \mu)/4 \Rightarrow \mu = 13$.

Q.2) The standard score for N(10, 9), value 7 is ?

Ans.) $z = (7 - 10)/3 = -1$

Q.3) For a huge data set how will you infer whether it follows normal distribution without plotting ?

Knowledge Check

Q.1) The standard score of value 25 is 3. If the variance is 16 then the mean value is ?

Ans.) $3 = (25 - \mu)/4 \Rightarrow \mu = 13$.

Q.2) The standard score for N(10, 9), value 7 is ?

Ans.) $z = (7 - 10)/3 = -1$

Q.3) For a huge data set how will you infer whether it follows normal distribution without plotting ?

Ans.) For a normal distribution mean = median = mode.

Hypothesis Testing

Hypothesis Testing

- Statistical hypothesis tests are based a statement called the **null hypothesis** that assumes nothing interesting is going on between whatever variables you are testing.
- The exact form of the null hypothesis varies from one type test to another: if you are testing whether groups differ, the null hypothesis states that the groups are the same.
- For instance, if you wanted to test whether the average age of voters in your home state differs from the national average, the null hypothesis would be that there is **no difference** between the average ages.

Why Null Hypothesis

- The purpose of a hypothesis test is to determine whether the null hypothesis is likely to be true given sample data.
- If there is little **evidence** against the null hypothesis given the data, you **accept** the null hypothesis.
- If the null hypothesis is **unlikely given the data**, you might **reject** the null in favor of the alternative hypothesis: that something interesting is going on.

The Alternative Hypothesis

The exact form of the alternative hypothesis will depend on the specific test you are carrying out.

If you are trying to determine if the average age of voters in your state and in the country are different,

- **Null Hypothesis**: There is no difference in the age in the two groups
- **Alternative Hypothesis**: the average age of voters in your state does in fact differ from the national average

p-value

Once you have the null and alternative hypothesis in hand, you choose a significance level (often denoted by the Greek letter **α**).

- The significance level is a probability threshold that determines when you reject the null hypothesis.
- After carrying out a test, if the probability of getting a result as extreme as the one you observe due to chance is lower than the significance level, you reject the null hypothesis in favor of the alternative.
- This probability of seeing a result as extreme or more extreme than the one observed is known as the **p-value**.

Interpreting the p-Value

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 (H_0)** the null hypothesis, so you **reject the null hypothesis.**

- A l**arge *p*-value (> 0.05)** indicates **weak evidence against (H_1)** the null hypothesis, so you **fail to reject the null hypothesis.**

- *p*-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the *p*-value so your readers can draw their own conclusions.

t-distribution

The T-test is a statistical test used to determine whether a numeric data sample of differs significantly from the population or whether two samples differ from one another.

One-Sample t-test

A one-sample t-test checks whether a sample mean differs from the population mean.

Two-Sample t-test

- A two-sample t-test investigates whether the means of two independent data samples differ from one another.

- In a two-sample test, the null hypothesis is that the means of both groups are the same.

- Unlike the one sample-test where we test against a known population parameter, the two sample test only involves sample means.

Paired t-test

- The basic two sample t-test is designed for testing differences between independent groups.
- In some cases, you might be interested in testing differences between samples of the same group at different points in time.
- For instance, a hospital might want to test whether a weight-loss drug works by checking the weights of the same group patients before and after treatment.
- A paired t-test lets you check whether the means of samples from the same group differ.

Type I and Type II error

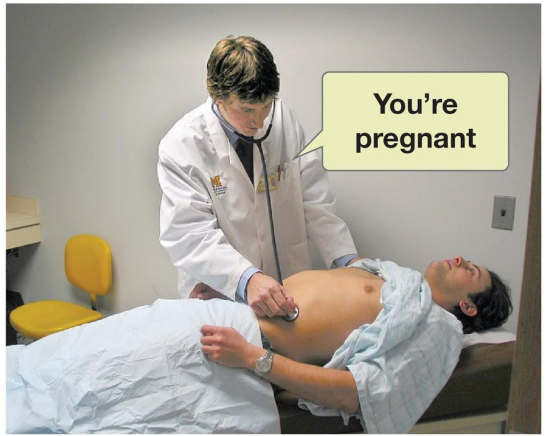
- The result of a statistical hypothesis test and the corresponding decision of whether to reject or accept the null hypothesis is not infallible.

- A test provides evidence for or against the null hypothesis and then you decide whether to accept or reject it based on that evidence, but the evidence may lack the strength to arrive at the correct conclusion.

- Incorrect conclusions made from hypothesis tests fall in one of two categories: **type I error** and **type II error**.

Errors in Hypothesis Testing

Type I error
(false positive)



Type II error
(false negative)

Type I error: We reject null hypothesis when it actually is true.

Type II error: We accept null hypothesis when it actually is false.

Type I error

Type I error describes a situation where you reject the null hypothesis when it is actually true.

- This type of error is also known as a **false positive** or false hit.

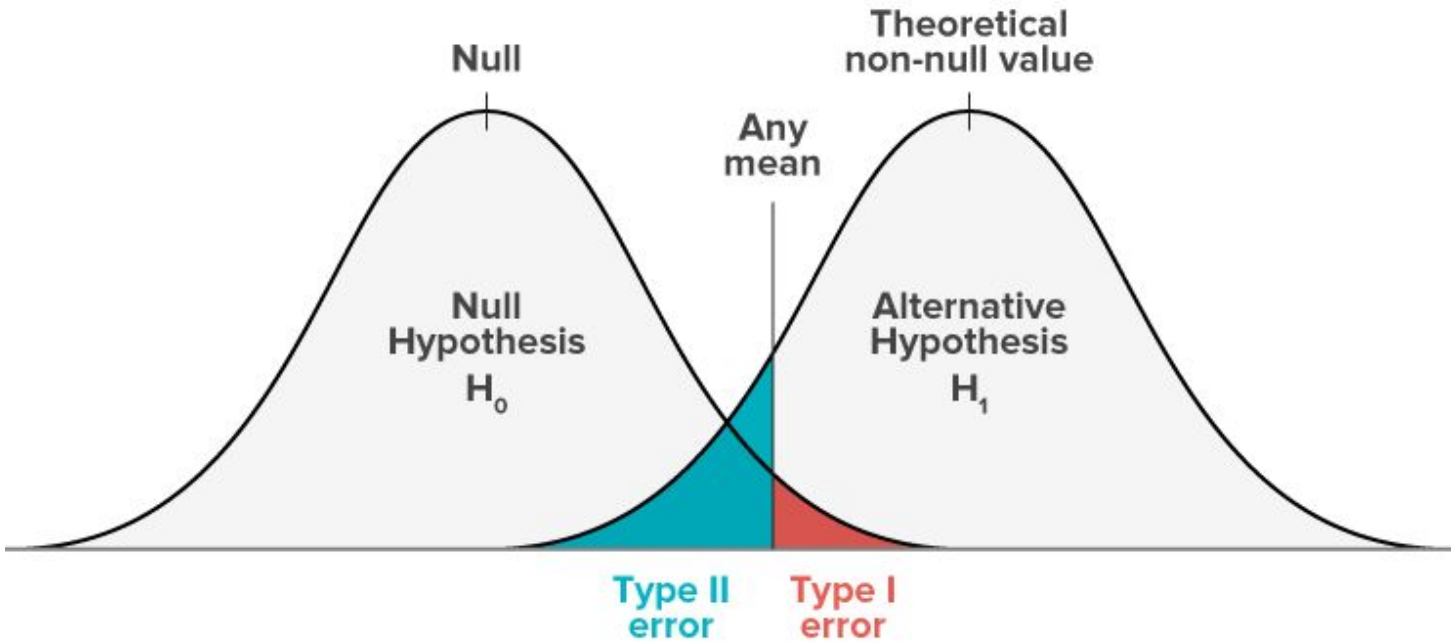
- The **type 1 error** rate is equal to the significance level **α**, so setting a higher confidence level (and therefore lower alpha) reduces the chances of getting a false positive.

Type II error

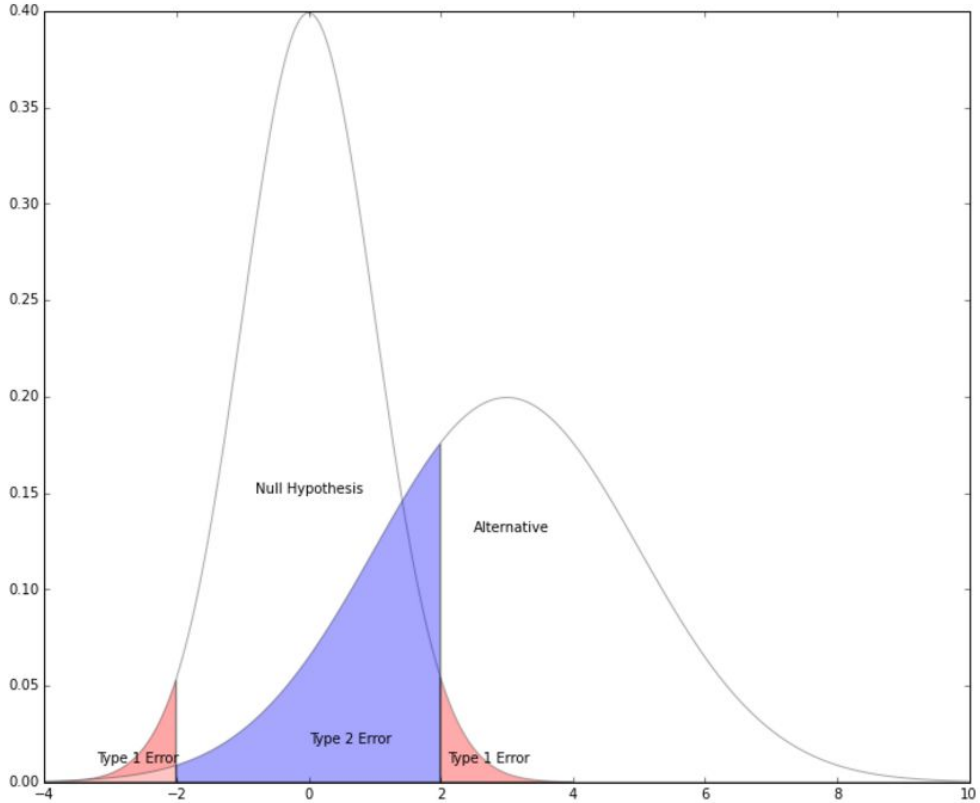
Type II error describes a situation where you fail to reject the null hypothesis when it is actually false.

- **Type II error** is also known as a **false negative** or miss. The higher your confidence level, the more likely you are to make a type II error.

Type I vs Type II error

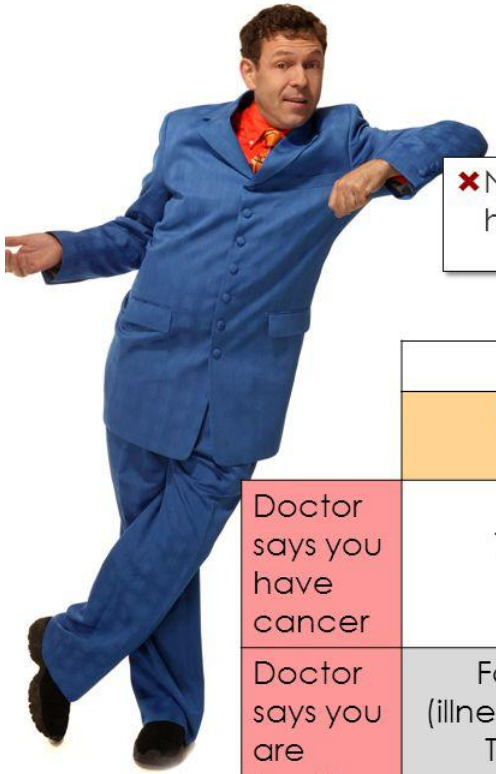


Type I vs Type II error



- The red areas indicate type I errors assuming the alternative hypothesis is not different from the null for a two-sided test with a 95% confidence level.
- The blue area represents type II errors that occur when the alternative hypothesis is different from the null, as shown by the distribution on the right.
- Note that the Type II error rate is the area under the alternative distribution within the quantiles determined by the null distribution and the confidence level.

Probabilities of Type I and Type II Errors



Which way would you lean?

✖ Null hypothesis: This person is not ill i.e. is healthy. There is nothing going on.

	Truth	
	Has cancer	No cancer i.e. healthy
Doctor says you have cancer	True positive	False positive (illness reported) TYPE 1 ERROR
Doctor says you are healthy	False negative (illness not detected) TYPE 2 ERROR	True negative

P(Type I error) = α

P(Type II error) = β

Type I vs Type II error

It is sometimes difficult to remember which error is which.

This infographics will help you remember!

Chi-Squared Goodness-Of-Fit Test

- We introduced the one-way t-test to check whether a sample mean differs from the an expected (population) mean.

- The chi-squared goodness-of-fit test is an **analog of the one-way t-test for categorical variables**: it tests whether the distribution of sample categorical data matches an expected distribution.

Chi-Squared Goodness-Of-Fit Test

- When working with categorical data the values the observations themselves aren't of much use for statistical testing because categories like "male", "female," and "other" have no mathematical meaning.

- Tests dealing with categorical variables are based on variable counts instead of the actual value of the variables themselves.

Chi-Squared Goodness-Of-Fit Test

For example, you could use a chi-squared goodness-of-fit test to check

- whether the race demographics of members at your church or school match that of the entire U.S. population, or,

- whether the computer browser preferences of your friends match those of Internet uses as a whole.

One-way ANOVA

- The one-way ANOVA tests whether the mean of some numeric variable differs across the levels of one categorical variable.

- It essentially answers the question: do any of the group means differ from one another?

The End