**Estatística Inferencial**

Estatísticas é uma das habilidades chaves fundamentais requeridos para deep learning. Qualquer profissional em deep learning certamente recomendaria aprender e aprimorar em estatísticas.

1. Porquê precisamos de estatística inferencial?

Suponhamos que você quer saber a média do salário de profissionais de Data Science na Índia. Quais dos métodos podem ser usados para calcular?

* 1. Encontrar todo profissional de data Science na Índia. Anotar os seus salários e calcular a média?
  2. Ou escolher um número de profissionais em uma cidade como Gurgaon. Anotar os seus salários e calcular a média.

Bem, o primeiro método não é impossível mas requer uma enorme quantidade de recursos e tempo. Mas hoje, empresas querem tomar decisões rápidas e de forma rentável, então o primeiro método não tem chance.

De outro modo, o segundo método parece possível. Mas, há um problema. Se a população de Gurgaon não for a mesma população da India inteira? Tem boas chances de você estimar errado o salário de Indianos profissionais em Data Science.

Agora, qual método pode ser usado para estimar a média de salário de todos os profissionais em Data Science na India?

Em termos simples, estatísticas inferenciais podem ser usadas para designar inferências além da data imediata disponível.

Com a ajuda de estatísticas inferenciais, nós podemos responder as questões abaixo:

* Extrair inferências sobre a população a partir de um exemplo.
* Concluir se um exemplo é significante diferente da população. Por exemplo, vamos dizer que você colecionou os detalhes de um salário de profissionais de Data Science em Bangalore. E você observou que a média de salário de profissionais de Data Science em Bangalore é maior que a média de salário na Índia. Agora, podemos concluir se a diferença é estaticamente significante.
* Se adicionando ou removendo uma característica de um modelo vai realmente ajudar a melhorar o modelo.
* Se um modelo é significativamente melhor que o outro?
* Testando hipótese em geral.

1. Pré-requisitos

Para começar com Estatísticas Inferenciais, é necessário ter um conhecimento em Estatísticas Descritivas. Você pode ver por [este curso](https://www.youtube.com/playlist?list=PLAwxTw4SYaPn22DmaF6x8JtG4TeWOJk_1) no youtube para aprender sobre estatísticas descritivas.

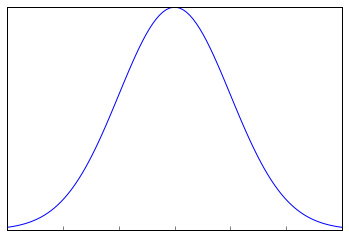
Aqui está uma breve descrição de algumas terminologias que vamos estar usando:

* **Estatística –** Uma única medida de algum atributo de um exemplo. Por exemplo: Média/Mediana/Modo de um exemplo de Cientistas de Dados em Bangalore
* **Estatística Populacional –** A estatística da população inteira em contexto. Por exemplo: Média populacional do salário de toda a população de cientistas de dados pela Índia.
* **Estatística de Amostra –** A estatística de um grupo retirado de uma população. Por exemplo: Média dos salários de todos os cientistas de dados em Bangalore.
* **Desvio padrão –** É a variação total de um dado populacional. É dado por σ.
* **Erro padrão –** É a variação total em um dado de exemplo. É relacionado ao desvio padrão como σ/√n, onde n, é o tamanho de exemplo.

1. Distribuição Amostral e Teorema do Limite Central

Suponhamos, que você anote o salário de 100 cientistas de dados em Gurgaon, calcule a média e repita o mesmo processo 200 vezes (arbitrariamente).

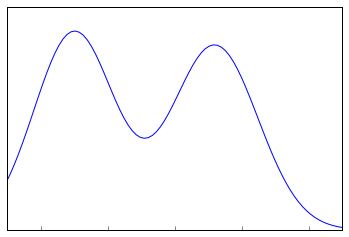
Quando você traça um gráfico de frequência dessas 200 vezes, você provavelmente vai ter uma curva como essa abaixo:



Essa curva parece muito com o que você viu em Estatísticas Descritivas. Isso é chamado de Distribuição Amostral ou o gráfico obteve esse resultado por traçar exemplos de média. Vamos olhar sobre uma descrição mais formal sobre Distribuição Amostral.

Uma distribuição amostral é uma distribuição de probabilidade de uma estatística obtida por meio de um grande números de exemplos desenhados de uma específica população.

Uma distribuição amostral se comporta muito como uma curva normal e possui algumas propriedades interessantes como:

* A forma da distribuição amostral não revela nada sobre a forma da população. Por exemplo, para a distribuição amostral acima, a distribuição da população pode ter resultados como esse gráfico abaixo: 

Distribuição Populacional

* Distribuição amostral ajuda a estimar a estatística populacional.

Mas como?

Isso vai ser explicado usando um teorema muito importante em estatísticas – **O Teorema do Limite Central**

* 1. Teorema do Limite Central

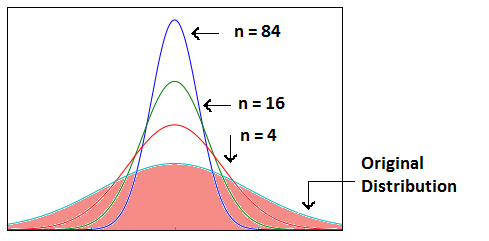
Afirma que quando é desenhado uma distribuição amostral de média, a média da amostra significa que vai ser igual a média populacional. E a distribuição amostral vai abordar uma distribuição normal com variância igual à σ/√n onde σ é o desvio padrão da população e n é o tamanho da amostra.

Pontos à notar:

* + 1. Teorema do Limite Central possui o verdadeiro resultado independente do tipo de distribuição da população.
    2. Agora, nós temos um jeito para estimar uma média populacional fazendo várias observações de exemplos de um tamanho fixo.
    3. Quanto maior o tamanho do exemplo, menor o erro padrão e maior a precisão em determinar a média populacional de uma média de exemplo.

Isso parece muito técnico, não é? Vamos ir a fundo para entender isso ponto por ponto.

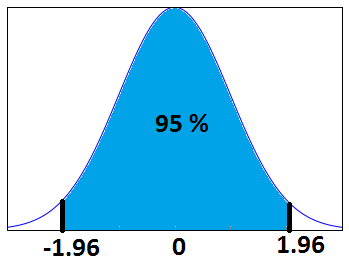
* Isso significa que – Não importa o tamanho da distribuição populacional, seja ela bimodal, enviesada à direita, etc. O tamanho da Distribuição Amostral vai continuar o mesmo (lembre-se da curva normal – formato de sino). Isso nos dá uma vantagem matemática para estimar a estatística populacional – não importa o tamanho da população.
* O número de amostras tem que ser suficientemente (geralmente mais que 50) para satisfatoriamente conquistar uma distribuição de uma curva normal. Também, é necessário um cuidado para manter o tamanho do exemplo fixado desde que qualquer mudança no tamanho da amostra vai mudar o tamanho da distribuição amostral e deixará de ser ‘formato de sino’.
* Conforme aumentamos o tamanho amostral, a distribuição amostral comprime dos dois lados dando para nós uma estimativa da estatística populacional melhor desde que se encontra em algum lugar no meio da distribuição amostral (geralmente). A imagem abaixo vai ajudar à você visualizar o efeito do tamanho amostral na forma da distribuição.



Agora, desde que nós coletamos os exemplos e colocamos as médias, é importante saber onde a média populacional se encontra de acordo com uma média amostral específica e o quão confiante pode ser. Isso nos trás o nosso próximo tópico – **Intervalo de confiança.**

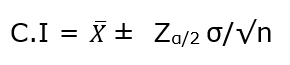
* 1. Intervalo de confiança

O intervalo de confiança é um tipo de intervalo estimado de uma distribuição amostral que dá um alcance dos valores onde a estatística populacional pode se encontrar. Vamos entender isso com um exemplo:



Nós sabemos que 95% dos valores encontram-se em 2 (1.96 para ser mais exato) desvios de padrões de uma distribuição normal de curva. Então, para a curva acima, a parte azul representa o intervalo de confianças de uma média amostral de 0.

Formalmente, Intervalo de Confiança é definido como:



Onde https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30112021/x-bar.pngé a amostra mediana

https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30111043/image_21.png = Valor de Z para o nível de confiança desejado α.

**σ** = Desvio padrão populacional

Para um valor de alfa de 0.95 i.e 95% intervalo de confiança, z=1.96.

Agora existe um termo em que você deve se familiarizar, **Margem de Erro**. É dada como {(z.σ)/√n} e é definida exemplificando o erro pela supervisão da pessoa que coletou os exemplos. Isso significa, se uma média amostral se encontrar no nível da margem de erro

* Existem intervalos de confiança diferentes para diferentes médias amostrais. Por exemplo, uma média amostral de 40 vai ter uma diferença de intervalo de confiança de uma média amostral de 45.
* Por 95% de intervalo de Confiança, nós não queremos dizer isso – a probabilidade de uma média populacional se encontrar em um intervalo é de 95%. Ao invés disso, 95% de Intervalo de Confiança significa que 95% do intervalo estimado vai conter a estatística populacional.

Muitas pessoas não tem um certo conhecimento sobre o intervalo de confiança e é frequentemente interpretado incorretamente. Então, eu gostaria que você pegasse um tempo estudando o 4º argumento e se aprofundasse.

* 1. **Exemplo prático**

Calcule 95% do intervalo de confiança para uma média amostral de 40 e um desvio padrão médio de 40 com um tamanho de exemplo igual à 100.

**Solução:**

Nós sabemos, valor-z para 95% do Intervalo de Confiança é 1.96. Consequentemente, Intervalo de confiança é calculado como:

C.I= [{x(bar) – (z\*s/√n)},{x(bar) – (z\*s/√n)}]

C.I = [{40-(1.96\*40/10},{ 40+(1.96\*40/10)}]

C.I = [32.16, 47.84]

1. **Testando Hipóteses**

Antes de eu entrar na explicação teórica, vamos entender Teste de Hipóteses usando um simples exemplo.

**Exemplo:** A 8ª Classe tem uma pontuação média de 40 de um total de 100 pontos. O diretor da escola decidiu que classes extras são necessárias em ordem para melhorar a performance daquela sala. A classe pontuou uma média de 45 pontos do total de 100 depois de pegar classes extras. Podemos ter certeza onde o aumento de pontos é um resultado das classes extras ou é randômico?

Teste de Hipóteses nos ajuda a descobrir isso. Faz uma estatística de exemplo ser verificada com uma estatística populacional ou uma estatística de outro exemplo para estudar qualquer intervenção e etc. Classes extras começam a intervenção no exemplo abaixo:

Teste de Hipóteses é definido em dois termos – **Hipóteses Nulas** e **Hipóteses Alternativas.**

* **Hipóteses Nula** faz a estatística de amostra ser igual a estatística populacional. Por exemplo: A Hipótese Nula para o exemplo abaixo iria ser que a média de pontos depois das classes extras é a mesma que antes da inclusão das classes extras..
* **Hipóteses Alternativas** para esse exemplo abaixo seria que os pontos depois das extras classes são significativamente diferentes antes das classes extras.

Teste de Hipóteses é feita em níveis diferentes de confiança e faz o uso do ponto-z para calcular a probabilidade. Então para 95% de Intervalo de Confiança, qualquer coisa abaixo do limiar de z para 95% vai rejeitar a Hipótese Nula.

Pontos para se anotar:

1. Nós não podemos aceitar a Hipótese Nula, somente rejeitá-la ou falhar em rejeitá-la
2. Como um tipo prático, Hipóteses Nula é geralmente mantida no que a gente quer refutar. Por exemplo: Você quer provar que estudantes aprendem melhor depois de assistirem aulas extras para a prova. A Hipótese Nula, nesse caso, vai ser que os pontos foram obtidos depois das aulas como o mesmo antes das aulas.