IM029 - Estatística

1. GER- História da Estatística

O termo estatística surge da expressão em latim *statisticum collegium*, palestra sobre os assuntos do Estado, de onde surgiu a palavra em língua italiana *statista*, que significa "homem de estado", ou político, e a palavra alemã *Statistik*, designando a análise de dados sobre o Estado.

A palavra foi proposta pela primeira vez no século XVII, em latim, por Schmeitzel, na Universidade de Jena e adotada pelo acadêmico alemão Godofredo Achenwall. Aparece como vocabulário na Enciclopédia Britânica em 1797, e adquiriu um significado de coleta e classificação de dados, no início do século XIX.

De acordo com a Revista do Instituto Internacional de Estatística, "Cinco homens, já receberam a honra de serem chamados de fundadores da estatística por diferentes autores":

- Gottfried Achenwall,
- John Graunt,
- Hermann Conring,
- Johann Peter Süssmilch e
- William Petty

Alguns autores dizem que é comum encontrar como marco inicial da estatística a publicação do *"Observations on the Bills of Mortality*" (*Observações sobre os Censos de Mortalidade, 1662*) de John Graunt.

Um fato conhecido é que as primeiras aplicações do pensamento estatístico estavam voltadas para as necessidades de Estado, na formulação de políticas públicas, fornecendo dados demográficos e econômicos. E a abrangência da estatística aumentou no começo do século XIX para incluir a análise de dados de maneira geral. Hoje, a estatística é largamente aplicada nas ciências naturais, e sociais, inclusive na administração pública e privada.

Os fundamentos matemáticos da estatística foram postos no século XVII com o desenvolvimento da teoria das probabilidades por Pascal e Fermat, que surgiu com o estudo dos jogos de azar.

O método dos mínimos quadrados foi descrito pela primeira vez por Carl Friedrich Gauss em 1794.

Com o avanço da computação e da tecnologia, a estatística experimentou um enorme crescimento. Métodos como análise de dados em grande escala, aprendizado de máquina (machine learning) e inteligência artificial se tornaram áreas proeminentes de pesquisa e aplicação.

Quando se ouve alguma notícia ou se lê algum artigo, muitas vezes números são mencionados;

- Exemplo 1: O número de carros vendidos no país aumentou em 30%
- Exemplo 2: A taxa de desempregos atinge hoje 7,5%;
- Exemplo 3: As ações da Empresa X subiram hoje R$1,5;
- Resultados do carnaval: 145 mortos e 2430 feridos;

Considerando um panorama histórico:

- Toda a ciência tem suas raízes na história do homem;
- A Matemática originou-se do convívio social, das trocas, da contagem, com característica prática utilitária e empírica, e é tida como "A Ciência que une a clareza do raciocínio à síntese da linguagem".
- A Estatística é o ramo da Matemática que teve origem semelhante;
- Desde a antiguidade vários povos já registravam o número de habitantes, de nascimentos, de óbitos, faziam estimativas de riquezas individuais e sociais, etc.
- Na idade média as informações eram colhidas geralmente com finalidade tributária;
- A partir do século XVI começaram a surgir as primeiras análises de fatos sociais, como batizados, casamentos, funerais, originando as primeiras tábuas e os primeiros números relativos;
- No século XVII o estudo de tais fatos foi adquirindo proporções verdadeiramente científicas;
- Godofredo Achenwall, batizou a nova ciência (ou método) com o nome de ESTATÍSTICA, determinando assim o seu objetivo e suas relações com a ciência.

Estatística, usada no singular pode ser é um número:

- Exemplo 4: No fechamento da bolsa, as ações da Vale foram cotadas a R$ 45,50;
- Exemplo 5: O total das vendas de uma empresa em um mês;

Estatísticas, (plural) é uma coleção de números:

- Exemplo 6: As vendas da Empresa Y totalizaram: 2,5 milhões em janeiro, 2,7 em fevereiro e 3.1 em março.

No entanto, o termo Estatística tem um sentido muito mais amplo do que apenas números, coleção de números, cálculos de medidas e traçado de gráficos. Assim, estatística pode ser definida como:

A ciência de coletar, organizar, apresentar, analisar e interpretar dados numéricos, para que conclusões possam ser extraídas deles, possibilitando a tomada de melhores decisões.

Como será mostrado mais adiante, a estatística se divide em:

- Descritiva: Envolve a descrição e resumo de dados. Isso inclui o uso de medidas de tendência central (como média, mediana e moda), medidas de dispersão (como desvio padrão e variância) e representações gráficas (como histogramas e gráficos de dispersão) para proporcionar uma visão geral dos dados. Aqui, os resultados apresentados não trazem inferências sobre um conjunto maior do que aquele que foi estudado. Ou seja, ela não se ocupa em aprender algo sobre uma população a partir de amostras extraídas dela. Por exemplo, a partir do estudo de características de uma classe de alunos, de um determinado colégio, não é o objetivo dela, fazer inferências sobre todos os alunos do colégio, com relação a essas características.
- Inferencial: Trata da inferência ou generalização de conclusões a partir de uma amostra para a população maior da qual a amostra foi retirada. Isso envolve a aplicação de testes de hipóteses, intervalos de confiança e métodos de estimativa para fazer afirmações sobre características desconhecidas da população com base em dados amostrais.

2. GER-Diferença entre Fenômenos Determinísticos e Estatísticos

2.1. Fenômenos Determinísticos

Os fenômenos determinísticos são eventos, ou processos, cujos resultados podem ser determinados por leis que governam as suas ocorrências ou por condições iniciais específicas. Ou seja, são previsíveis e podem ser totalmente compreendidos se conhecemos as leis que os regem e as condições do sistema, aos quais se associam. Com exemplos, podemos citar o movimento de planetas no sistema solar de acordo com as leis da gravidade, o comportamento de um pêndulo simples e a solução de equações matemáticas.

2.2. Fenômenos Estatísticos

Os fenômenos estatísticos são eventos, ou processos, que exibem variabilidade e aleatoriedade, ou seja, apresentam resultados que não podem ser previstos de maneira exata. No entanto, esses fenômenos podem ser modelados e analisados através de métodos estatísticos para compreender padrões ou tendências. Por exemplo, resultados de jogos de azar, flutuações no mercado de ações e a variação no tempo de chegada de clientes a um serviço.

Sejam determinísticos ou estatísticos, a abordagem dos fenômenos exige a construção de modelos matemáticos, cuja validade deve ser comprovada por dados de observação, ou resultados dos experimentos, que permitirão comparar o que foi obtido com o que era esperado.

2.3. Modelos Determinísticos

Em um modelo determinístico, as condições sob as quais um experimento é executado determinem o seu resultado. Por exemplo, se introduzirmos uma bateria em um circuito simples, o modelo matemático, que descreve a corrente que o percorre, é $I = E/R$, isto é, a lei de Ohm. Com os valores de E e R dados, é possível calcular o valor de I. E esse modelo seria aplicável para a maioria das aplicações, independentes dos pequenos desvios que viessem a ocorrer, provocados, p. ex. por temperatura e umidade. Por depender apenas das condições sob as quais o experimento ou o procedimento é executado, o modelo determinístico, permite que se chegue a um resultado efetivo.

2.4. Modelos Estatísticos, não Determinísticos ou Probabilísticos

Mais à frente será mostrado como esses modelos podem ser apresentados. Mas por hora, vamos exemplificar uma situação em que ele deve ser usado, já que é impossível usar uma abordagem determinística. Trata-se de um caso em se deseja determinar a precipitação da chuva decorrente de uma tempestade, que cairá em um determinado local. Para isso, dispõe-se de (1) instrumentos para registrar a precipitação e de (2) dados da observação meteorológicas sobre a tempestade que se avizinha: pressão barométrica em vários pontos, variações de pressão, velocidade do vento, origem e direção da tormenta, e várias leituras referentes a altitudes elevadas. Contudo, ainda que essas informações sejam valiosas para um prognóstico da natureza geral da precipitação (fraca, média ou forte), elas não possibilitam determinar quanta chuva irá cair. Ou seja, estamos diante de um fenômeno que não se presta a um tratamento determinístico. Um modelo probabilístico explica a situação de forma melhor. Em um modelo não-determinístico, as condições de experimentação determinam somente o comportamento probabilístico (mais especificamente, a lei probabilística) do resultado. Em outras palavras, enquanto que, em um modelo determinístico o conhecimento dos valores de um conjunto de parâmetros nos possibilita prever um resultado, em um modelo probabilístico por mais que conheçamos valores de parâmetros relacionados ao evento não será possível prever um resultado específico. Para estes tipos de eventos, como será visto mais adiante, é possível apenas definir uma distribuição de probabilidade, e através dela, calcular a probabilidade de ocorrer um valor ou outro de resultado, dentro de uma determinada faixa, ou intervalo, de valores.

2.5. Diferenças entre Fenômenos Estatísticos e Determinísticos

- Aleatoriedade vs. Determinismo: Nos fenômenos estatísticos, a aleatoriedade desempenha um papel significativo, e os resultados não podem ser previstos com certeza. Em contraste, os fenômenos determinísticos são governados por regras e condições iniciais específicas, tornando-os previsíveis.
- Modelagem Estatística vs. Modelagem Matemática: Fenômenos estatísticos são frequentemente modelados usando distribuições de probabilidade e métodos estatísticos. Fenômenos determinísticos são modelados usando equações matemáticas precisas que descrevem o comportamento determinístico do sistema.
- Incerteza vs. Certezas: A incerteza é uma característica intrínseca dos fenômenos estatísticos, enquanto os fenômenos determinísticos geralmente são associados a certezas e previsibilidade.
- Variação vs. Consistência: Fenômenos estatísticos exibem variação entre as observações, enquanto fenômenos determinísticos são consistentes, com resultados idênticos sob as mesmas condições.

2.6. Exemplos de experimentos não determinísticos

Estamos agora em condições de examinar o que entendemos por um experimento "aleatório" ou "não-determinístico".

- E1: Jogue um dado e observe o número mostrado na face de cima.
- E2: Jogue uma moeda quatro vezes e observe o número de caras obtido.
- E3: Jogue uma moeda quatro vezes e observe a sequência obtida de caras e coroas.
- E4: Em uma linha de produção, fabrique peças em série e conte o número de peças defeituosas produzidas em um período de 24 horas.
- E5: Uma asa de avião é fixada por um grande número de rebites. Conte o número de rebites defeituosos.
- E6: Uma lâmpada é fabricada. Em seguida é ensaiada quanto à duração da vida, pela colocação em um soquete e anotação do tempo decorrido (em horas) até queimar.
- E7: Um lote de 10 peças contém 3 defeituosas. As peças são retiradas uma a uma (sem reposição da peça retirada) até que a

última peça defeituosa seja encontrada. O número total de peças retiradas do lote é contado.

- E8: Peças são fabricadas até que 10 peças perfeitas sejam produzidas. O número total de peças fabricadas é contado.
- E9: Um míssil é lançado. Em um momento especificado t, suas três velocidades componentes, V_x, V_y e V_z, são observadas.
- E10: Um míssil recém lançado é observado nos instantes t_1, t_2, •••, t_n. Em cada um desses instantes, a altura do míssil acima do solo é registrada.
- E11: A resistência à tração de uma barra metálica é medida.
- E12: De uma urna, que só contém bolas pretas, tira-se uma bola e verifica-se sua cor.
- E13: Um termógrafo registra a temperatura continuamente, num período de 24 horas. Em determinada localidade e em uma data especificada, esse termógrafo é lido.
- E14: Na situação descrita em E18, x e y, 118 temperaturas mínima e máxima, no período de 24 horas considerado, são registradas.

O que os experimentos acima têm em comum?

(a) Cada experimento poderá ser repetido indefinidamente sob condições essencialmente inalteradas.

(b) Embora não sejamos capazes de afirmar que resultado particular ocorrerá, seremos capazes de descrever todos os possíveis resultados do experimento;

(c) Quando o experimento for executado algumas vezes os resultados individuais parecerão ocorrer de forma acidental. Mas quando executado um grande número de vezes, uma configuração definida ou regularidade surgirá. É esta regularidade que torna possível construir um modelo matemático preciso com o qual se analisará o experimento. Isso vale tanto para o caso do lançamento de moedas que iremos mencionar a seguir bem como para todos os experimentos acima. Nas repetidas jogadas de uma moeda equilibrada, muito embora caras e cordas apareçam sucessivamente, em uma maneira quase arbitrária, a prática nos mostra que depois de um grande número de jogadas o número de caras e coroas é aproximadamente o mesmo. Para que um experimento fique bem definido, é necessário saber como ele deve ser feito e qual característica numérica dele (entre as várias que ele eventualmente pode fornecer) estamos interessados em observar (caso de E2 e E3).

3. GER - Divisão da Estatística em Descritiva e Indutiva (ou Inferencial)

A estatística se divide basicamente em duas partes, a Descritiva e a Inferencial. São duas abordagens distintas usadas na análise de dados em diferentes contextos. Em resumo, a estatística descritiva é usada quando o objetivo é descrever e resumir características fundamentais de um conjunto de dados específico, enquanto a estatística inferencial é aplicada quando se deseja fazer inferências sobre uma população com base em uma amostra, como testar hipóteses ou fazer previsões.

3.1. Estatística descritiva

É a parte da estatística que lida com a organização, resumo e apresentação de dados numéricos com o objetivo de transformá-los em informação.

Estatística descritiva, exemplos de uso:

- Exemplo D1 - Pesquisas de opinião em grupos pequenos de pessoas: Após conduzir uma pesquisa de opinião com 500 entrevistados sobre a preferência de sabores de sorvete, a estatística descritiva seria usada para calcular a média, mediana e moda dos sabores preferidos, apresentando uma visão geral das preferências da amostra;
- Exemplo D2 - Avaliação de Desempenho Acadêmico: Ao analisar as notas de uma turma em uma disciplina específica, a estatística descritiva seria aplicada para calcular a média das notas, a variabilidade e identificar possíveis padrões de desempenho;
- Exemplo D3 -Análise de Vendas Mensais: Uma empresa deseja entender melhor seu desempenho de vendas mensais. A estatística descritiva seria utilizada para calcular a média de vendas, a variação e visualizar esses dados por meio de gráficos.

3.2. Estatística Indutiva ou Inferencial

É a parte da estatística que, baseando-se em resultados obtidos da análise de uma amostra da população, procura inferir, induzir ou estimar as leis de comportamento da população da qual a amostra foi retirada. Neste contexto, o significado de população é bem amplo, referindo-se ao conjunto de todos os elementos que se deseja estudar, ou seja, uma coleção de todos os possíveis elementos, objetos ou medidas de interesse. A estatística indutiva existe porque, muitas vezes, apesar dos recursos computacionais e da boa vontade, não é possível estudar todo um conjunto de dados de interesse. Neste caso, estuda-se uma parte do conjunto. O principal motivo para se trabalhar com uma parte do conjunto, ao invés do conjunto inteiro é o custo. É através da estatística indutiva que podemos aceitar ou rejeitar hipóteses que podem surgir sobre as características da população, a partir, também, da análise da amostra representativa dessa população. Exemplos:

- O conjunto das rendas de todos os habitantes de Campinas;
- O conjunto de todas as notas dos alunos de Estatística;
- O conjunto das alturas de todos os alunos da Universidade.

Um levantamento efetuado sobre toda uma população é levantamento censitário ou simplesmente censo. Fazer um censo sobre toda uma população é muito difícil, caro e leva muito tempo. Então, o que se faz é trabalhar com partes da população denominadas amostras, que são uma porção ou parte de uma população de interesse. Utilizar amostras para se ter conhecimento sobre populações é realizado intensamente na Agricultura, Política, Negócios, Marketing, Governo, etc., como se pode ver pelos seguintes exemplos.

- Estatística inferencial ou indutiva, exemplos de uso:
    - Exemplo I1 - Testes de Hipóteses em Experimentos Científicos: Um pesquisador realiza um experimento para testar um novo medicamento. A estatística inferencial seria usada para realizar testes de hipóteses e determinar se há evidências estatísticas significativas de que o medicamento é eficaz;
    - Exemplo I2 - Previsão de Eleições: Antes de uma eleição, por exemplo, diversos órgãos de pesquisa e imprensa ouvem um conjunto selecionado de eleitores para ter uma ideia do desempenho dos vários candidatos nas futuras eleições. Com base em pesquisas de opinião de uma amostra representativa, a estatística inferencial seria aplicada para

fazer inferências sobre a intenção de voto da população como um todo, usando intervalos de confiança e estimativas;

- Exemplo I3 - Controle de Qualidade em Manufatura: Uma fábrica produz um determinado componente e deseja inferir a qualidade média da produção com base em uma amostra. Em intervalos de tempo toma uma amostra desse componente para verificar se o processo está sob controle e evitar a fabricação de itens defeituosos. A estatística inferencial seria usada para calcular intervalos de confiança para a média da qualidade do componente;
- Exemplo I4: O IBGE faz levantamentos periódicos sobre emprego, desemprego, inflação, etc;
- Exemplo I5: Redes de rádio e TV se utilizam constantemente dos índices de popularidade dos programas para fixar valores da propaganda ou então modificar ou eliminar programas com audiência insatisfatória;
- Exemplo I6: Biólogos marcam pássaros, peixes, etc. para tentar prever e estudar seus hábitos;

3.3. Amostragem na Estatística Indutiva

É o processo de escolha de uma amostra da população. É um processo que envolve riscos, pois toma-se decisões sobre toda a população com base em apenas uma parte dela.

3.4. A Teoria da Probabilidade na Estatística Indutiva

É utilizada para fornecer uma ideia do risco envolvido, ou seja, do erro que se comete ao utilizar uma amostra ao invés de toda a população, desde que, é claro, a amostra seja selecionada através de critérios probabilísticos, isto é, ao acaso.

3.5. Definição mais Completa de Estatística Indutiva

Baseado nos conceitos anteriores pode-se definir Estatística Indutiva ou Inferencial como: a coleção de métodos e técnicas utilizado para se estudar uma população, baseados em amostras probabilísticas desta mesma população, sendo que as principais técnicas utilizadas são:

- Estimação: visa determinar o valor dos parâmetros desconhecidos.
- Testes de hipóteses: visa testar suposições acerca das características de uma certa população

4. DESC - Estatística Descritiva - Introdução

A estatística descritiva, como já vimos, é o ramo da estatística que se preocupa com a descrição e resumo de características importantes de conjuntos de dados.

Importância da análise descritiva na compreensão dos dados

A estatística descritiva permite, através das ações mencionadas abaixo, explorar, entender e comunicar informações importantes contidas nos dados estatísticos, fornecendo uma base sólida para análises mais avançadas e tomadas de decisão fundamentadas.

Ações da estatística descritiva

- Coleta de Dados:

  O primeiro passo é reunir os dados relevantes para a análise. Esses dados podem ser coletados por meio de pesquisas, experimentos, observações ou fontes existentes.

- Organização dos Dados:

  A análise descritiva resume os dados de forma concisa e compreensível, permitindo uma visão geral rápida e clara das características principais do conjunto de dados, como tendências centrais, dispersão, forma da distribuição, presença de valores atípicos, entre outros. Isso pode envolver a criação de tabelas de frequência, listagens ou outras representações que destaquem a estrutura do conjunto de dados.

- Apresentação dos Dados:

  A estatística descritiva utiliza gráficos e diagramas para apresentar visualmente as características dos dados. Histogramas, gráficos de dispersão, boxplots e outras representações gráficas são comumente empregados.

- Identificação de padrões:

  Ao examinar estatísticas descritivas como média, mediana, moda, desvio padrão, quartis, etc., é possível identificar padrões nos dados, como tendências, sazonalidades, flutuações e comportamentos incomuns.

- Detecção de outliers:

A análise descritiva permite identificar valores atípicos ou outliers que podem distorcer as análises estatísticas. Isso é crucial para entender se esses valores são erros de medição, anomalias reais nos dados ou pontos de dados significativos que exigem atenção especial.

- Medidas de Tendência Central:

São calculadas medidas que indicam onde o centro do conjunto de dados está localizado. Isso inclui a média, a mediana e a moda.

- Medidas de Dispersão:

São calculadas medidas que indicam a extensão ou dispersão dos dados. O desvio padrão, a variância e a amplitude são exemplos de medidas de dispersão.

- Medidas de Posição:

Percentis e quartis são utilizados para dividir o conjunto de dados em partes específicas, fornecendo informações sobre a posição relativa dos valores.

- Comparação entre grupos:

Ao comparar grupos diferentes de dados, a análise descritiva ajuda a identificar diferenças e semelhanças nas características dos grupos, destacando insights importantes sobre variabilidade, centralidade e distribuição.

- Comunicação eficaz:

As estatísticas descritivas fornecem uma linguagem comum para descrever e comunicar informações sobre os dados. Os resultados obtidos são comunicados de forma clara e concisa. Isso pode envolver relatórios, apresentações gráficas ou outras formas de comunicação, dependendo do público-alvo.
Isso é essencial para garantir que as conclusões sejam compreendidas de forma clara e precisa por todos os interessados, independentemente do nível de conhecimento estatístico.

- Interpretação e Conclusões:

Com base nas medidas calculadas e nas representações visuais, são feitas interpretações e conclusões sobre as características centrais, a dispersão e a distribuição dos dados.

- Tomada de decisão informada:

Uma compreensão sólida dos dados por meio da análise descritiva é essencial para embasar a tomada de decisões informadas em uma variedade de contextos, incluindo negócios, ciência, saúde, governo, entre outros.

A seguir, estão descritos os principais passos realizados dentro da estatística descritiva.

4.1. Coletas de Dados

O primeiro passo é coletar os dados relevantes para a análise. Isso pode envolver a realização de pesquisas, a coleta de dados de fontes secundárias ou a obtenção de dados experimentais.

4.2. Organizar os dados

Os dados coletados precisam ser organizados de maneira adequada para análise. Isso pode incluir a tabulação dos dados em tabelas, a criação de gráficos ou outros métodos de organização.

4.3. Representação gráfica dos dados

A representação gráfica dos dados é uma parte importante da estatística descritiva. Gráficos como histogramas, gráficos de barras, gráficos de dispersão e box plots são frequentemente utilizados para visualizar os dados e identificar padrões ou tendências.

4.4. Medidas de tendência central

As medidas de tendência central são utilizadas para descrever onde os dados tendem a se concentrar. As medidas mais comuns incluem a média, a mediana e a moda.

4.5. Medidas de dispersão

As medidas de dispersão indicam a variabilidade ou espalhamento dos dados em torno de uma medida central. Exemplos de medidas de dispersão incluem o desvio padrão, a variância e a amplitude interquartil.

4.6. Análise de distribuição

A distribuição dos dados pode ser examinada para determinar se eles seguem uma distribuição específica, como a distribuição normal. Isso pode ser feito visualmente por meio de gráficos ou estatisticamente por meio de testes de normalidade.

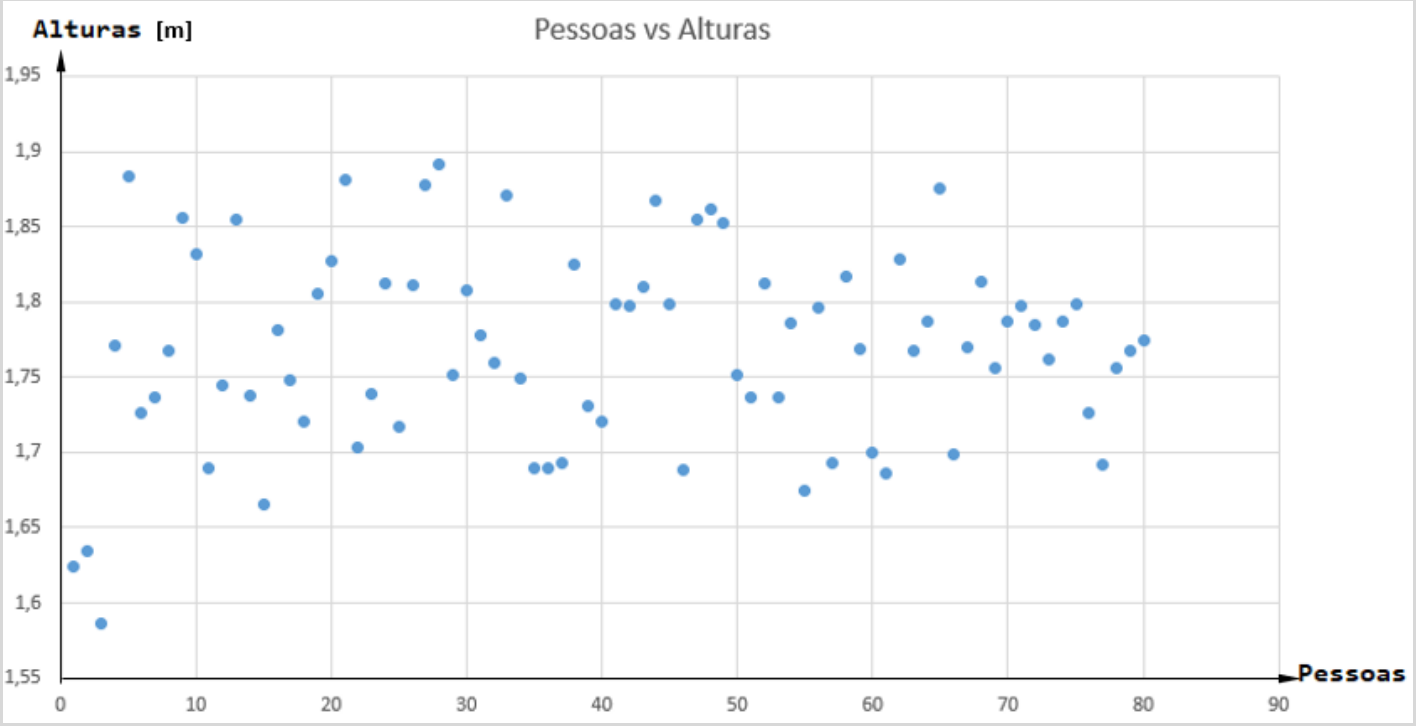
5. DESC - Coleta e Organização dos Dados

Vamos iniciar pelo seguinte exemplo. Em um grupo de 80 pessoas do sexo masculino, foi feito um levantamento das alturas.

Pessoa	Altura	Pessoa	Altura	Pessoa	Altura	Pessoa	Altura
1	1,624	21	1,881	41	1,799	61	1,686
2	1,634	22	1,703	42	1,797	62	1,828
3	1,586	23	1,739	43	1,81	63	1,768
4	1,771	24	1,812	44	1,867	64	1,787
5	1,883	25	1,717	45	1,798	65	1,875
6	1,726	26	1,811	46	1,688	66	1,699
7	1,736	27	1,878	47	1,855	67	1,77
8	1,768	28	1,891	48	1,862	68	1,813
9	1,856	29	1,752	49	1,852	69	1,756
10	1,832	30	1,808	50	1,752	70	1,787
11	1,69	31	1,778	51	1,737	71	1,797
12	1,744	32	1,76	52	1,812	72	1,785
13	1,855	33	1,871	53	1,737	73	1,762
14	1,738	34	1,749	54	1,786	74	1,787
15	1,665	35	1,689	55	1,674	75	1,799
16	1,781	36	1,69	56	1,796	76	1,726
17	1,748	37	1,693	57	1,693	77	1,692
18	1,721	38	1,825	58	1,817	78	1,756
19	1,805	39	1,731	59	1,769	79	1,768
20	1,827	40	1,721	60	1,7	80	1,774

Olhando para a tabela acima, não é possível ver, rapidamente, quais são as tendências. Qual é altura que aparece com mais frequência, quais são os extremos, qual é o valor médio, etc.

Se colocarmos esses valores em um gráfico de Altura vs Pessoas, seremos capazes de ver os valores extremos, mas não muito mais que isso.


Alturas [m]
Pessoas vs Alturas
1,95
1,9
1,85
1,8
1,75
1,7
1,65
1,6
1,55
0
10
20
30
40
50
60
70
80
90
Pessoas

6. DESC - Representação Gráfica dos Dados

Existem maneiras melhores de apresentar os dados coletados, de forma que algumas informações possam ser extraídas mais facilmente. São elas:

- Tabela de Frequências;
- Gráfico de Frequências;
- Polígono de Frequências;

6.1. Tabela de Frequências

A tabela de frequências permite categorizar os dados levantados, porque nela, são definidos vários intervalos de valores (com limite inferior e superior), ou classes, e em cada uma dessas classes são colocados o número de indivíduos levantados que se encaixam nelas. No caso do exemplo, são colocados o número de pessoas cujas alturas se encaixam dentro de cada um dos intervalos de valores de altura ou em cada classe de altura.

Em outras palavras, a tabela de frequências, da forma como é montada, disponibiliza um acesso rápido ao número absoluto e ao percentual do número de elementos observados que pertencem a cada classe definida.

Esta tabela de frequências é montada da seguinte forma.

- Identificar no conjunto de dados levantados, quais são o maior e o menor valor. No caso do exemplo acima, o menor é **1,586m** e o maior é **1,891m**;
- Com esses dois valores, calcular a diferença entre o maior e o menor valor. Essa diferença é denominada amplitude (**a**) dos valores levantados. No caso, **a** = **1,891m** - **1,586m** = **0,305m**;
- Dividir a amplitude (**a**) em vários intervalos (mínimo 5 e máximo 20). No caso do exemplo, o número de intervalos escolhido, denominados classes, foi 6. Existe uma regra empírica!! Cada intervalo, ou classe, terá um limite inferior e um limite superior, sendo que, o limite superior da classe atual é o limite inferior da próxima;
- Colocar as 6 classes em 6 linhas da primeira coluna da tabela de frequências (coluna classes);
- Na segunda coluna da tabela, colocar, em cada classe, a quantidade de pessoas cujas alturas se encaixam nela (coluna de frequências absolutas);

- Na terceira coluna da tabela, colocar, em cada classe, o percentual de pessoas cujas alturas se encaixam nela (coluna de frequências relativas);
- Na quarta coluna de tabela, colocar, em cada classe, a soma das frequências absolutas da classe atual e das anteriores (coluna de frequências absolutas acumuladas);
- Na quinta coluna de tabela, colocar, em cada classe, a soma das frequências relativas da classe atual e das anteriores (coluna de frequências relativas acumuladas);

Classes [m]	Frequências Absolutas (Pessoas em cada Classe)	Frequências Relativas (Percentual de Pessoas em cada Classe)	Frequências Absolutas Acumuladas	Frequências Relativas Acumuladas
1,586 a 1,637	3	0,0375	3	0,0375
1,637 a 1,688	4	0,0500	7	0,0875
1,688 a 1,739	20	0,2500	27	0,3375
1,739 a 1,789	23	0,2875	50	0,6250
1,789 a 1,840	18	0,2250	68	0,8500
1,840 a 1,891	12	0,1500	80	1,0000

6.2. Histograma

Em um histograma, temos uma sequência de vários retângulos colados uns nos outros, montados da seguinte maneira.

No eixo x são colocadas as classes (em número > = 5 e <= 20), as mesmas definidas para a montagem da tabela de frequências, cujos tamanhos, representando as bases dos retângulos.

No eixo y são colocados os valores das alturas dos retângulos. Estes valores representam o número de ocorrências (absoluto ou percentual) por "tamanho da classe" (base do retângulo), ou seja, o número de ocorrências verificado dentro da classe. Então, essa razão (número de ocorrências)/(tamanho da classe) é uma medida de densidade, ou seja, as alturas dos retângulos no eixo y do histograma são medidas de densidade, semelhantemente ao que ocorre com os valores do eixo y no caso do gráfico de densidade de probabilidade, como veremos mais adiante.

No gráfico do histograma os únicos pares (x, y), que fazem sentido, são:

- os que estão no eixo x, pares (x, 0), onde x são os valores que delimitam os tamanhos das classes. Os valores específicos de x que estão presentes dentro da tabela de dados levantadas, são importantes apenas para sabermos a que classe pertencem, mas uma vez situados dentro dela, esses valores não assumem nenhum significado especial dentro do histograma;
- os que estão no eixo y, pares (0, y), onde y são os valores de densidade (número de elementos/tamanho da classe) das várias classes. Os demais valores pertencentes ao eixo y, também não possuem um significado

Para calcular o número de elementos em uma classe, número absoluto ou percentual, basta calcular a área do retângulo associado à classe. A área será a altura do retângulo multiplicada pela sua base:

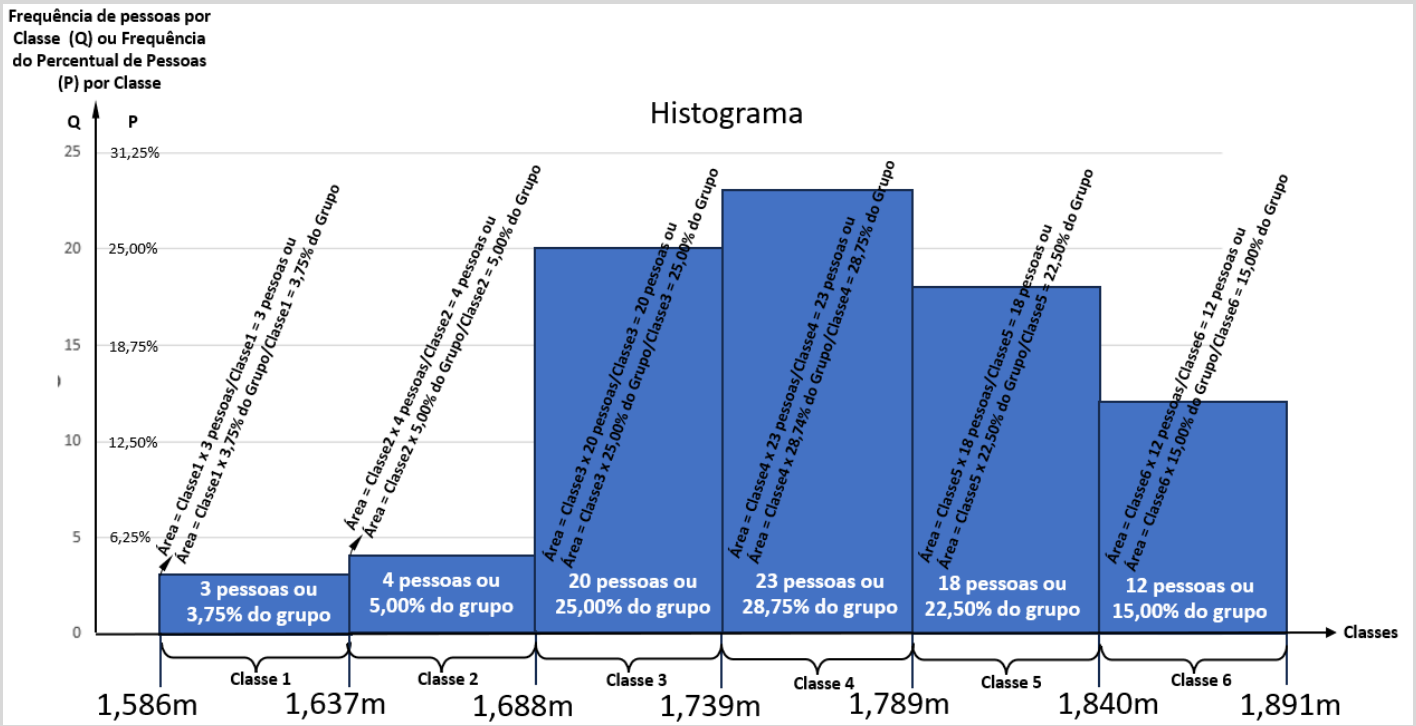
Número de elementos em uma classe = Área do retângulo.

Área do retângulo = Base x Altura.

Base = Tamanho da Classe

Altura = Densidade de Elementos por Tamanho da Classe = (Número de elementos) / (Tamanho da Classe)

Número de elementos na classe = (Tamanho da Classe)x(Densidade de Elementos da Classe)



O Histograma fornece informações melhores do que a tabela com os dados coletados.

Agora, há um fato muito interessante pode ser notado no Histograma. Mais para frente, no estudo das variáveis aleatórias contínuas, irá aparecer a função densidade de probabilidade (fdp) relacionadas às variáveis aleatórias contínuas.

Neste tipo de função, as probabilidades são calculadas em trechos (e não em pontos específicos do eixo x como é possível fazer com as variáveis discretas nas funções de probabilidade) determinando-se a área sob a curva da fdp no trecho. E, na maioria dos livros textos, não existem explicações detalhadas sobre como se chega à equação destas funções (como, p. ex., à equação da fdp gaussiana ou normal). Mas sem tentar explicar como as fdps são elaboradas, é possível fazer um paralelo entre elas o Histograma.

Vamos considerar o seguinte. No eixo **y** dos Histograma, se denominarmos os valores colocados lá como "densidade de pessoas por classe" ao invés de "frequência de pessoas por classe", não estaremos propondo nada de errado, uma vez que as unidades usadas nas duas denominações são as mesmas, [número de pessoas/Classe].

Feita esta primeira consideração, poderíamos fazer uma segunda, que seria estabelecer um paralelo entre Histogramas e funções de densidade de probabilidade.

Uma das características de uma fdp, é que a integral dela vale 1, quando calculado de – infinito a + infinito. No Histograma, não estaríamos fazendo nada de errado se, ao invés de trabalharmos com valores percentuais de 0 a 100%, trabalharmos com valores de décimos da unidade de 0 a 1. P. ex., na Classe 1, poderíamos expressar a frequência (ou densidade) com o valor 0,0375 ao invés de 3,75%. Dessa forma, podemos considerar que, no Histograma, a integral, calculada de – infinito a + infinito, vale 1, também. E embora no exemplo apresentado não estejamos mencionando probabilidades diretamente, nada nos impede de formular a seguinte questão. Se escolhêssemos uma pessoa, ao acaso, dentro desse grupo, qual seria a probabilidade dela pertencer, p. ex., à Classe 1? A resposta é 0,0375 ou 3,75%.

Essa relação entre Histogramas e fdps, dá uma pequena pista de qual poderia ter sido o ponto de partida, usado pelos matemáticos antigos, para obtê-las.

6.3. Polígono de Frequências (completar!!)

O polígono de frequências é um gráfico utilizado na estatística para representar visualmente a distribuição de frequências de um conjunto de dados. Ele é construído a partir de um histograma, onde os pontos médios de cada classe de intervalo são plotados no eixo horizontal e as frequências de cada classe são plotadas no eixo vertical.

A partir do polígono de frequências, várias informações podem ser extraídas:

(1) Forma da distribuição: O polígono de frequências pode indicar a forma geral da distribuição dos dados. Por exemplo, se a linha do polígono é aproximadamente reta, os dados podem ter uma distribuição uniforme; se a linha é inclinada para cima ou para baixo, os dados podem ter uma distribuição crescente ou decrescente, respectivamente; se a linha tem picos ou depressões, os dados podem ter uma distribuição irregular.

(2) Centralidade: O ponto médio do polígono de frequências pode ser utilizado como uma medida da centralidade dos dados. Pelo polígono de freqências é possível verificar rapidamente o valor médio que ocorre em cada uma das classes do histograma.

(3) Dispersão: A dispersão dos dados pode ser avaliada observando a amplitude horizontal do polígono de frequências. Quanto maior a dispersão, mais ampla será a base do polígono.

(4) Outliers: Os outliers, ou valores atípicos, podem ser identificados visualmente como pontos que estão distantes do padrão geral do polígono de frequências.

(4) Tendências: Padrões ou tendências nos dados podem ser identificados através do comportamento geral da linha do polígono de frequências. Por exemplo, se a linha está subindo ou descendo constantemente, isso pode indicar uma tendência nos dados ao longo do tempo ou de outra variável relevante.

(5) Comparação: O polígono de frequências também pode ser utilizado para comparar diferentes conjuntos de dados. Plotando os polígonos de frequências lado a lado, é possível visualizar diferenças nas distribuições, centralidades, dispersões, etc.

(**COLOCAR FIGURA!!**)

6.4. Gráficos de Barras (completar!!)

6.5. Gráficos de Setores (completar!!)

6.6. Tabelas de contingência e frequência cruzada (completar!!)

7. DESC - Estatística Descritiva – Medidas

Seguem as principais medidas usadas pela estatística descritiva para descrever as características principais de um conjunto de dados.

- Medidas de Tendência Central:
  - Média: É a média aritmética dos valores em um conjunto de dados.
  - Mediana: É o valor que divide o conjunto de dados ao meio quando eles estão organizados em ordem crescente.
  - Moda: É o valor que ocorre com mais frequência no conjunto de dados.

- Medidas de Dispersão:
  - Desvio Padrão: Mede a dispersão dos valores em relação à média.
  - Amplitude: É a diferença entre o maior e o menor valor no conjunto de dados.
  - Variância: É o quadrado do desvio padrão.

- Medidas de Posição:
  - Percentis: Indicam a porcentagem de dados que está abaixo de um determinado valor.
  - Quartis: Dividem o conjunto de dados em quatro partes iguais.

- Visualização de Dados:
  - Histogramas: Gráficos de barras que representam a distribuição de frequência de um conjunto de dados.
  - Gráficos de dispersão: Mostram a relação entre duas variáveis.
  - Boxplots (Diagramas de Caixa): Representam visualmente a distribuição estatística dos dados.

- Medidas de Associação:

- Correlação: Avalia a força e a direção da relação linear entre duas variáveis.
- Coeficiente de Determinação (R^2): Indica a proporção da variabilidade em uma variável que pode ser explicada pela outra em modelos de regressão.

- Tabelas de Frequência:

  Apresentam a contagem ou a porcentagem de observações em diferentes categorias.

  A aplicação dessas metodologias depende do tipo de dados disponíveis e dos objetivos específicos da análise. A estatística descritiva é fundamental para resumir e compreender as características essenciais de conjuntos de dados, fornecendo insights valiosos para a tomada de decisões.

8. DESC - Medidas de Tendência Central

8.1. Média Aritmética

A média aritmética, mais comumente conhecida como "média", é uma medida de tendência calculada pela soma de uma lista de números dividida pelo número de itens. A média é útil para determinar a tendência geral de um conjunto de dados fornecendo o que seria um valor típico dele. A vantagem da média é que ela é muito fácil de ser calculada. No entanto, se tomada sozinha, em um conjunto de dados com um alto número de outliers ou com uma distribuição muito irregular, a média deixa de ser um bom parâmetro para evidenciar uma tendência e consequentemente, para embasar alguma tomada de decisão.

8.2. Mediana

É uma medida de tendência expressa pelo valor que divide o conjunto de dados ao meio. Quando temos:

- um número ímpar de dados, a mediana é o valor na posição (n-1)/2 + 1
- um número par de dados, a mediana é a média entre os valores n/2 e n/2 + 1

8.3. Moda

É uma medida de tendência expressa pelo valor que aparece o maior número de vezes dentro do conjunto de dados.

8.4. Média Ponderada

A média ponderada é frequentemente utilizada em levantamentos estatísticos quando diferentes observações têm importâncias relativas diferentes e é necessário calcular uma média que leve isso em consideração, atribuindo pesos maiores aos itens mais importantes.

Exemplos:

- Pesquisas de opinião: Quando uma pesquisa de opinião é realizada e certas respostas são consideradas mais representativas ou relevantes do que outras, essas respostas podem receber um peso maior ao calcular a média.

- Avaliação de desempenho: Em avaliações de desempenho, como avaliações de produtos, serviços ou desempenho de funcionários, certos aspectos podem ser considerados mais críticos do que outros. Nesse caso, pesos diferentes são atribuídos aos diferentes aspectos ao calcular a média ponderada.

- Notas escolares: Em sistemas de avaliação educacional, como em escolas e universidades, diferentes disciplinas ou tarefas podem ter pesos diferentes. Por exemplo, uma prova final pode valer mais do que trabalhos em grupo ou atividades de sala de aula. Ao calcular a média final, esses pesos são levados em consideração.

- Índices financeiros: Em análises financeiras, como o cálculo de índices de preços ou taxas de retorno, diferentes componentes podem ter pesos diferentes. Por exemplo, ao calcular um índice de preços ao consumidor, itens mais importantes no orçamento familiar podem receber pesos maiores.

9. DESC - Medidas de Dispersão

9.1. Amplitude, ou Intervalo ou Dispersão

É a diferença entre o maior e o menor valor, existentes dentro do conjunto de dados;

9.2. Variância

É uma medida da dispersão dos dados em torno da média. Em outras palavras, a variância mede o quanto os valores individuais de um conjunto de dados variam em relação à média desse conjunto. Uma variância elevada indica uma dispersão maior dos dados em torno da média, enquanto uma variância pequena indica uma dispersão menor e, portanto, uma maior concentração dos dados em torno da média.

Existem dois tipos de variância:

- **Variância da População:** é usada quando temos dados que representam toda a população que estamos estudando e
- **Variância da Amostra:** é usada quando temos apenas uma amostra dos dados da população e queremos fazer inferências sobre ela.

A raiz quadrada da variância da população é o **desvio padrão da população** e a raiz quadrada da variância da amostra é o **desvio padrão da amostra**. O desvio padrão é outra medida comumente usada para avaliar a dispersão dos dados

1. **Variância da População:**
   - A variância da população é usada quando temos dados que representam toda a população que estamos estudando.
   - A fórmula para calcular a variância da população é:
     $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$
   - Nesta fórmula, σ^2 representa a variância populacional, N é o tamanho da população, x_i são os valores individuais da população, e μ é a média da população.
2. **Variância da Amostra:**
   - A variância da amostra é usada quando temos apenas uma amostra dos dados da população e queremos fazer inferências sobre a população maior.
   - A fórmula para calcular a variância da amostra é:
     $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
   - Nesta fórmula, s^2 representa a variância da amostra, n é o tamanho da amostra, x_i são os valores individuais da amostra, e $\bar{x}$ é a média da amostra.

A diferença crucial entre as duas fórmulas está no denominador. Na fórmula da variância da população, dividimos pela quantidade total de observações N enquanto na fórmula da variância da amostra, dividimos pela diferença entre a quantidade de observações e 1 (N - 1).

Essa correção é chamada de correção de Bessel e é usada na variância da amostra para compensar a tendência da amostra de subestimar a variabilidade da população. Isso ajuda a produzir uma estimativa menos tendenciosa da variância populacional com base na amostra.

9.3. Desvio Padrão

É uma medida de dispersão que indica o quanto os valores de um conjunto de dados se desviam, em média, em relação à média desse conjunto. Em outras palavras, ele fornece uma medida da dispersão dos dados em torno da média. Um desvio padrão elevado indica uma maior dispersão dos dados, enquanto um desvio padrão pequeno indica uma dispersão menor e, portanto, uma maior concentração dos dados em torno da média.

Existem dois tipos de variância:

- **Desvio Padrão da População:** é a raiz quadrada da variância populacional e
- **Desvio Padrão da Amostra:** é a raiz quadrada da variância da amostra.

1. **Desvio Padrão da População:**
   - O desvio padrão da população é a raiz quadrada da variância populacional.
   - É calculado usando a fórmula:

     $$\sigma = \sqrt{\sigma^2}$$

   - Onde σ representa o desvio padrão da população e σ^2 representa a variância da população.

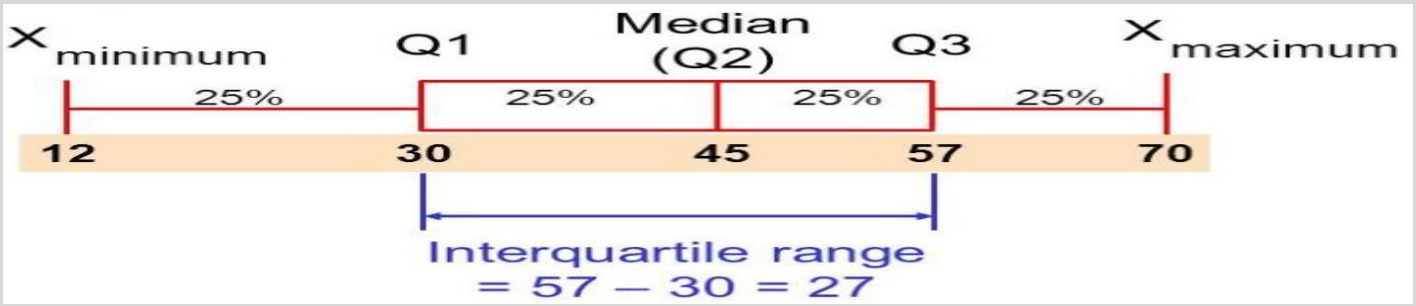
2. **Desvio Padrão da Amostra:**
   - O desvio padrão da amostra é a raiz quadrada da variância da amostra.
   - É calculado usando a fórmula:

     $$s = \sqrt{s^2}$$

   - Onde s representa o desvio padrão da amostra e s^2 representa a variância da amostra.

9.4. Quartis e Intervalo Inter Quartil

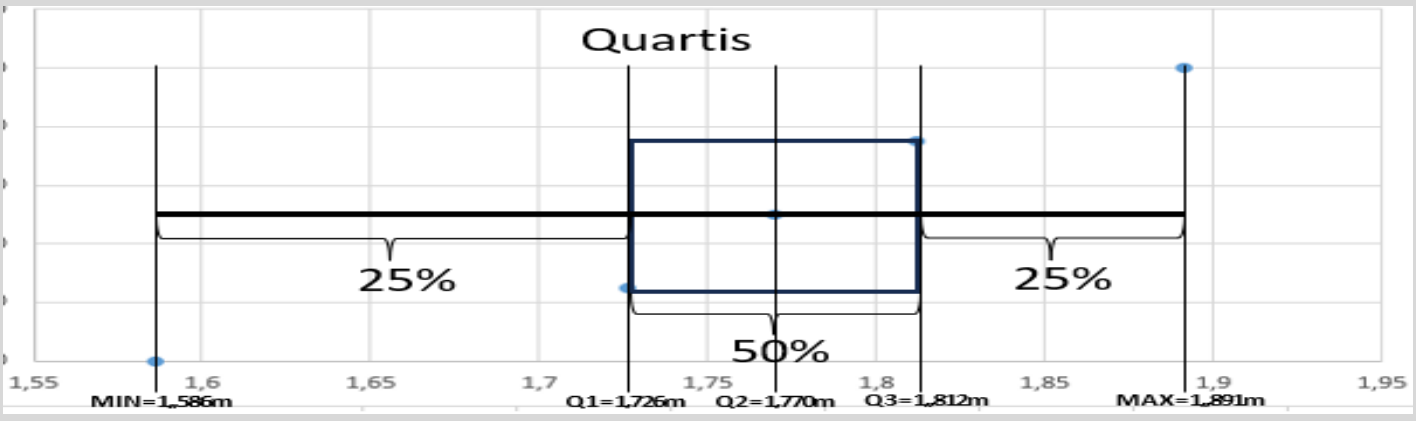
Os quartis e o intervalo interquartil são medidas úteis para descrever a dispersão e a centralização dos dados em um conjunto de observações, que ajudam a resumir e entender a distribuição desses dados.



- **Quartis**: os quartis são valores que dividem um conjunto de dados ordenados em quatro partes iguais, representando assim os 25%, 50% e 75% dos dados:
  - O primeiro quartil (Q1) é o valor abaixo do qual está o 25% inferior dos dados.
  - O segundo quartil (Q2), também conhecido como mediana, é o valor que divide o conjunto de dados em duas partes iguais, representando o 50% central dos dados.
  - O terceiro quartil (Q3) é o valor abaixo do qual está o 75% inferior dos dados.

  Para o caso do exemplo colocado no item 5, temos a tabela mostrada abaixo e o gráfico, na sequência.

Trecho		Casos
Valor MIN. a Q1	1,586m a 1,726m	25%
Q1 a Q2	1,726m a 1,770m	25%
Q2 a Q3	1,770m a 1,812m	25%
Q3 a valor MAX.	1,812m a 1,890m	25%



- **Intervalo interquartil (IQR)**: o intervalo interquartil é a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1), ou seja, IQR = Q3 - Q1. O IQR:
    - representa a amplitude dos 50% centrais dos dados e fornece uma medida de dispersão que é menos sensível a valores extremos do que a amplitude total dos dados;
    - é uma medida robusta de variabilidade que é frequentemente usada para identificar a dispersão dos dados em torno da mediana.

  Para o caso do exemplo colocado no item 5, o intervalo interquartil, IQR, vale IQR = Q3 - Q1 = **1,812m - 1,726m = 0,086m**.

10. DESC – Outras Representações Gráficas

10.1. Box plot. (completar!!)

10.2. Gráfico de dispersão (completar!!)

10.3. Gráfico de linha (completar!!)

10.4. Gráfico de densidade (completar!!)

11. DESC- Medidas de Assimetria e Curtose

11.1. Coeficiente de assimetria

O coeficiente de assimetria, também conhecido como coeficiente de Skewness, é uma medida estatística que descreve a assimetria da distribuição de dados em relação à sua média. Em outras palavras, ele indica se os dados estão distribuídos de forma simétrica ou assimétrica em torno da média.Se o coeficiente de assimetria for:

- zero, isso indica que a distribuição dos dados é perfeitamente simétrica.
- positivo, indica que a cauda direita da distribuição é mais longa do que a cauda esquerda, ou seja, a distribuição é assimétrica positiva ou à direita.
- negativo, indica que a cauda esquerda da distribuição é mais longa do que a cauda direita, ou seja, a distribuição é assimétrica negativa ou à esquerda.

O coeficiente de assimetria pode ser calculado de várias maneiras, mas uma das fórmulas mais comuns é usando momentos estatísticos. Aqui está a fórmula para o coeficiente de assimetria:

$$\text{Coeficiente de Assimetria} = \frac{n}{(n-1)(n-2)} \times \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{s^3}$$

Onde:

- n é o número de observações na amostra.
- x_i é cada valor na amostra.
- $\bar{x}$ é a média da amostra.
- s é o desvio padrão da amostra.

Essa medida é útil para entender a forma e a distribuição dos dados, o que pode ter implicações importantes na análise estatística e na tomada de decisões.

11.2. Coeficiente de curtose

O coeficiente de curtose, também conhecido como medida de achatamento, é uma medida estatística que descreve a forma da distribuição dos dados em relação à sua média. Ele indica o grau de afilamento ou achatamento das caudas de uma distribuição em relação à distribuição normal.Em resumo, o coeficiente de curtose fornece informações sobre a "espessura" das caudas da distribuição e pode ser útil na comparação de diferentes distribuições de dados. Assim como o coeficiente de assimetria, a curtose é uma medida importante para entender a forma e a distribuição dos dados em uma análise estatística.

Se o coeficiente de curtose for:

- zero, isso indica que a distribuição dos dados tem o mesmo achatamento que uma distribuição normal.
- positivo, indica que a distribuição tem caudas mais pesadas e é mais afilada que uma distribuição normal; isso é conhecido como curtose excessiva ou leptocurtica.
- negativo, indica que a distribuição é mais achatada e tem caudas mais leves que uma distribuição normal; isso é conhecido como curtose subexcessiva ou platicurtica.

O coeficiente de curtose pode ser calculado de várias maneiras, mas uma das fórmulas mais comuns é usando momentos estatísticos. Aqui está a fórmula para o coeficiente de curtose:

$$\text{Coeficiente de Curtose} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \times \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Onde:

- n é o número de observações na amostra.
- x_i é cada valor na amostra.
- $\bar{x}$ é a média da amostra.
- s é o desvio padrão da amostra.

11.3. Interpretação dos resultados (completar!!)

12. DESC – Análise Exploratória de Dados

12.1. Identificação de outliers

Identificar outliers em um conjunto de dados é uma etapa importante em um levantamento estatístico para garantir a qualidade e a precisão das análises. Existem várias técnicas que podem ser usadas para identificar outliers. No entanto, é importante lembrar que a identificação de outliers não é uma ciência exata e pode variar dependendo do contexto e dos objetivos da análise. É recomendável utilizar uma combinação de técnicas e julgamento humano para identificar e lidar com outliers de forma adequada.

Seguem algumas técnicas para identificar outliers:

(1) Visualização de dados: Gráficos como histogramas, box plots, scatter plots e gráficos de séries temporais podem fornecer insights visuais sobre a distribuição dos dados e destacar observações que estão fora do padrão;

(2) Método do intervalo interquartil (IQR): Este método envolve calcular o intervalo interquartil (IQR), que é a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1), e identificar outliers como valores que estão abaixo de Q1 - 1,5IQR e Q3 + 1,5IQR;

(3) Método z-score: Este método envolve calcular o z-score para cada observação, que indica quantos desvios padrão uma observação está da média. Observações com z-scores muito altos ou muito baixos (geralmente acima de 3 ou abaixo de -3) são consideradas outliers.

(4) Análise de resíduos: Em modelos estatísticos, como regressão linear, os resíduos (diferença entre os valores observados e os valores previstos) podem ser analisados para identificar outliers.

(5) Testes estatísticos: Testes estatísticos como o teste de Grubbs ou o teste de Dixon podem ser usados para identificar valores discrepantes em conjuntos de dados.

(6) Conhecimento de domínio: Em algumas situações, o conhecimento especializado sobre o domínio dos dados pode ajudar na identificação de outliers. Por exemplo, se os dados representam medidas físicas, valores extremamente altos ou baixos podem ser identificados com base no conhecimento das limitações físicas.

12.2. Análise de padrões e tendências

Análises de padrões e tendências são fundamentais em estatística para entender o comportamento dos dados ao longo do tempo ou em relação a outras variáveis. A escolha da técnica mais apropriada depende do tipo de dados, do contexto da análise e dos objetivos específicos da investigação.

Aqui estão algumas técnicas comuns para realizar essas análises:

- Gráficos de séries temporais: Plotar os dados ao longo do tempo em um gráfico de linha ou outro tipo de gráfico de séries temporais pode ajudar a identificar padrões e tendências visuais.

- Médias móveis: Calculando médias móveis (médias dos valores em um intervalo de tempo específico) pode suavizar flutuações e revelar tendências subjacentes nos dados.

- Análise de regressão: Utilizar técnicas de regressão para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes, possibilitando a identificação de padrões e tendências.

- Decomposição de séries temporais: Decompor a série temporal em componentes como tendência, sazonalidade e variação aleatória pode ajudar a entender os diferentes padrões presentes nos dados.

- Testes estatísticos: Utilizar testes estatísticos para avaliar se há uma tendência significativa nos dados ao longo do tempo, como o teste de Mann-Kendall ou o teste de Sen's Slope.

- Modelagem de séries temporais: Aplicar modelos estatísticos ou matemáticos específicos para prever ou explicar padrões e tendências nos dados ao longo do tempo, como modelos ARIMA (Autoregressive Integrated Moving Average) ou modelos de suavização exponencial.

- Análise de correlação: Avaliar a correlação entre variáveis pode ajudar a identificar padrões e tendências em relação a outras variáveis.

- Análise de sazonalidade: Identificar padrões sazonais nos dados pode revelar tendências que se repetem ao longo de determinados períodos do ano.

12.3. Detecção de relações entre variáveis

Para detectar a relação entre variáveis em estatística, você pode empregar diversas técnicas, dependendo do tipo de dados e da natureza da relação que você está tentando explorar. A escolha da técnica mais apropriada depende da natureza dos dados, do tipo de relação que você está tentando explorar e dos objetivos específicos da análise.

Segur uma relação das técnicas mais comuns:

- Análise de correlação: A correlação estatística mede a força e a direção da relação linear entre duas variáveis. O coeficiente de correlação de Pearson é frequentemente utilizado para variáveis contínuas, enquanto o coeficiente de correlação de Spearman é mais adequado para variáveis ordinais ou quando a relação não é linear.

- Gráficos de dispersão: Plotar os valores das duas variáveis em um gráfico de dispersão pode fornecer uma visualização direta da relação entre elas. Padrões visuais, como tendências lineares, não lineares ou agrupamentos, podem indicar diferentes tipos de relação entre as variáveis.

- Análise de regressão: A análise de regressão permite modelar a relação entre uma variável dependente e uma ou mais variáveis

independentes. A regressão linear é a forma mais simples de análise de regressão e é útil para investigar relações lineares entre as variáveis. Modelos de regressão mais complexos podem ser aplicados para capturar relações não lineares ou relações entre múltiplas variáveis independentes e uma variável dependente.

- Análise de covariância: A análise de covariância (ANCOVA) é uma extensão da análise de variância (ANOVA) que leva em consideração uma variável contínua adicional (chamada de covariável) que pode influenciar a variável dependente.

- Análise de séries temporais: Se as variáveis estão relacionadas ao longo do tempo, a análise de séries temporais pode ser usada para investigar a relação e identificar padrões temporais.

- Testes de hipóteses específicos: Dependendo da natureza da relação que você está interessado em investigar, você pode usar testes de hipóteses específicos, como testes t ou ANOVA, para comparar as médias de grupos ou subgrupos em relação às variáveis de interesse.

- Análise de componentes principais: Se você está interessado em encontrar relações entre várias variáveis independentes e uma variável dependente, a análise de componentes principais pode ajudar a reduzir a dimensionalidade dos dados e identificar os principais padrões de variação.

13. DESC – Transformação de Dados

13.1. Normalização

A normalização de dados é uma técnica comum em estatística usada para ajustar a escala dos dados para que eles possam ser comparados de maneira mais significativa. Existem várias maneiras de normalizar os dados, mas uma das técnicas mais comuns é a normalização min-max. No entanto, é importante lembrar que a normalização pode alterar a distribuição dos dados e, em certos casos, pode ser necessário considerar outras técnicas de pré-processamento de dados.

Aqui está como você pode realizar a normalização min-max:

Passo 1: Identifique a faixa de valores dos seus dados: Determine o valor mínimo (min) e o valor máximo (max) para cada variável que você deseja normalizar.

Passo 2: Aplique a fórmula de normalização min-max para cada valor nos seus dados: Para cada valor (xi) na variável que você deseja normalizar, aplique a seguinte fórmula:

$$x_i' = \frac{x_i - \min}{\max - \min}$$

Onde:

- x_i' é o valor normalizado de x_i.

Passo 3: Repita o processo para todas as variáveis que você deseja normalizar: Aplique a mesma fórmula para cada variável, utilizando o seu próprio mínimo e máximo.

Passo 4: Interprete os dados normalizados: Os valores normalizados estarão na faixa de 0 a 1, com 0 representando o valor mínimo original e 1 representando o valor máximo original. Valores entre 0 e 1 representam a proporção do valor original em relação à faixa total de valores.

13.2. Padronização

A padronização de dados, também conhecida como normalização z-score, é outra técnica comum em estatística utilizada para ajustar a escala dos dados. Ela é útil quando você deseja comparar diferentes variáveis que podem ter escalas diferentes ou quando deseja garantir que as variáveis tenham a mesma escala e influência em uma análise. Essa técnica é amplamente utilizada em várias aplicações estatísticas, como análise de regressão, análise de cluster e análise de componentes principais.

Aqui está como você pode realizar a padronização z-score:

Passo 1: Calcule a média e o desvio padrão dos seus dados: Para cada variável que você deseja padronizar, calcule a média (μ) e o desvio padrão (σ).

Passo 2: Aplique a fórmula de padronização z-score para cada valor nos seus dados: Para cada valor (xi) na variável que você deseja padronizar, aplique a seguinte fórmula:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Onde:

- z_i é o valor padronizado de x_i.
- μ é a média da variável.
- σ é o desvio padrão da variável.

Passo 3: Repita o processo para todas as variáveis que você deseja padronizar: Aplique a mesma fórmula para cada variável, utilizando a sua própria média e desvio padrão.

Passo 4: Interprete os dados padronizados: Os valores padronizados têm uma média de 0 e um desvio padrão de 1. Um valor padronizado de 0 indica que o valor está na média da distribuição original, enquanto valores positivos ou negativos indicam quantos desvios padrão o valor está da média.

13.3. Log-transformação

A transformação logarítmica é uma técnica comum em estatística utilizada para modificar a distribuição dos dados, especialmente quando os dados estão muito espalhados ou quando apresentam uma distribuição assimétrica. É importante notar que a transformação logarítmica só pode ser aplicada a dados estritamente positivos, já que o logaritmo de zero ou números negativos não é definido nos números reais. Além disso, a interpretação dos resultados após a transformação logarítmica pode diferir da interpretação dos dados originais, e os resultados devem ser interpretados com isso em mente.

Aqui está como você pode realizar a transformação de log:

Passo 1: Identifique os seus dados: Determine a variável ou conjunto de dados que você deseja transformar utilizando a transformação logarítmica.

Passo 2: Escolha a base do logaritmo: A transformação logarítmica pode ser realizada utilizando diferentes bases de logaritmo, como logaritmo natural (base e), logaritmo na base 10, ou logaritmo na base 2. A escolha da base depende do contexto da análise e da interpretação dos resultados desejados.

Passo 3: Aplique a função logarítmica para cada valor nos seus dados: Para cada valor (xi) na variável que você deseja transformar, aplique a função logarítmica com a base escolhida:

$$y_i = \log(x_i)$$

Onde:

- y_i é o valor transformado de x_i.

Passo 4: Interprete os dados transformados: A transformação logarítmica é especialmente útil para reduzir a assimetria e a heterocedasticidade dos dados, tornando-os mais adequados para determinadas análises estatísticas, como análise de regressão. Valores extremamente altos são "puxados para baixo" e valores extremamente baixos são "puxados para cima", resultando em uma distribuição mais uniforme dos dados.

14. DESC – Resumo Numérico e Gráfico para Diferentes Tipos de Variáveis

14.1. Estatísticas descritivas para variáveis categóricas

As variáveis categóricas são aquelas usadas para descrever características ou atributos, com um número pequeno de valores não. Como, por exemplo:

- gênero: masculino, feminino, outro;
- estado civil: casado, solteiro, viúvo, separado;
- tipo sanguíneo: A, B, AB ou O.
- tipo de uma rocha: ígnea, sedimentar ou metamórfica.

Os valores associados a essas variáveis não podem ser colocadas em ordem de magnitude, crescente ou decrescente. Ainda assim, podem ser ordenadas ou agrupadas de alguma maneira.

Para esse tipo de variável não faz sentido falar de média ou mediana. A única medida de tendência central cabível para ela é a moda. E embora exista a possiblidade de atribuir números às variáveis categóricas (rótulos), para facilitar o processamento estatístico, isso não implica na possibilidade de manipulá-los, e realizar com eles operações aritméticas como normalmente se faz com os números. Com elas só é possível fazer operações relacionadas a conjuntos.

As variáveis categóricas podem ser classificadas em dois tipos, as nominais e as ordinais.

- **Variáveis nominais**: São variáveis categóricas que não possuem uma ordem ou hierarquia natural entre as categorias. Exemplos incluem cor (vermelho, azul, verde), tipo de veículo (carro, caminhão, motocicleta) e gênero (masculino, feminino, outros). As categorias são mutuamente exclusivas e não há noção de ordenação entre elas.
- **Variáveis ordinais**: Ao contrário das variáveis nominais, as variáveis ordinais apresentam uma ordem ou hierarquia natural entre as categorias. Isso significa que, embora as variáveis ordinais representem grupos, esses grupos têm uma sequência ou classificação intrínseca. Exemplos incluem nível de educação (ensino fundamental, ensino médio, graduação, pós-graduação), faixas de renda (baixa, média, alta), e classificação de satisfação (insatisfeito, neutro, satisfeito). É importante notar que, embora as variáveis ordinais possam ser ordenadas, as diferenças entre as categorias

não são necessariamente uniformes ou mensuráveis. Por exemplo, a diferença entre "ensino fundamental" e "ensino médio" pode não ser a mesma que entre "graduação" e "pós-graduação", em termos de anos de estudo ou experiência de aprendizado.

Ao analisar dados categóricos, técnicas específicas de análise estatística são aplicadas, uma vez que métodos projetados para dados numéricos não são adequados para dados que representam categorias. O entendimento e o uso correto de variáveis categóricas são essenciais para a interpretação adequada de pesquisas e análises estatísticas em diversas áreas, incluindo ciências sociais, marketing, pesquisa de mercado, saúde pública, e muitas outras.

14.2. Estatísticas descritivas para variáveis contínuas (completar!!)

x

14.3. Considerações específicas para variáveis qualitativas e quantitativas (completar!!)

x

15. DESC – Aplicações e Interpretações

15.1. Exemplos práticos de aplicação dos conceitos definidos (completar!!)

x

15.2. Interpretação dos resultados da análise descritiva (completar!!)

x

16. DESC – Considerações Éticas e Limitações

16.1. Ética na análise de dados descritivos (completar!!)

x

16.2. Limitações da análise descritiva e possíveis viesses (completar!!)

x

17. PROB - Fundamentos da Teoria das Probabilidades:

A ideia deste texto é tentar chegar à base das definições e conceitos da estatística inferencial.

17.1. Espaços amostrais, eventos, frequência relativa e probabilidade

Espaço Amostral (S):

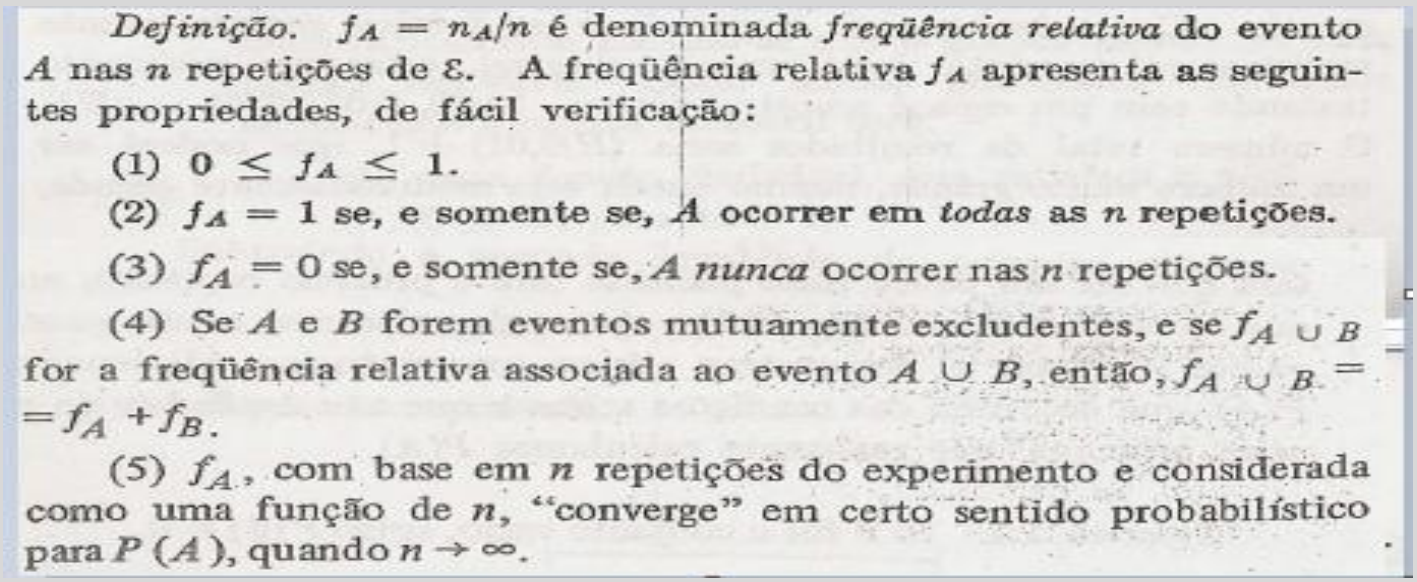
É o conjunto de todos os resultados possíveis de um experimento aleatório, ou seja, evento cujo resultado não pode ser definido com certeza. É denotado por S e pode ser finito, infinito contável ou infinito não contável. Por exemplo, ao lançar um dado, o espaço amostral é {1, 2, 3, 4, 5, 6}.

Evento (E):

É um subconjunto do espaço amostral, ou seja, é um conjunto de resultados possíveis. Um evento ocorre se o resultado do experimento está contido no conjunto de resultados que define o evento. Pode ser simples (um único resultado) ou composto (mais de um resultado). Exemplo: No lançamento do dado, o evento "sair um número par" pode ser representado por E = {2, 4, 6}.

Frequência relativa

Vamos supor que repetimos n vezes um experimento E e sejam A e B, dois eventos associados a E. Suponhamos que nA e nB sejam os números de vezes que o evento A e o evento B ocorreram, respectivamente, durante as n repetições.



Definição. $f_A = n_A/n$ é denominada *freqüência relativa* do evento A nas n repetições de $\mathcal{E}$. A freqüência relativa f_A apresenta as seguintes propriedades, de fácil verificação:

(1) $0 \leq f_A \leq 1$.

(2) $f_A = 1$ se, e somente se, A ocorrer em *todas* as n repetições.

(3) $f_A = 0$ se, e somente se, A *nunca* ocorrer nas n repetições.

(4) Se A e B forem eventos mutuamente excludentes, e se $f_{A \cup B}$ for a freqüência relativa associada ao evento $A \cup B$, então, $f_{A \cup B} = f_A + f_B$.

(5) f_A, com base em n repetições do experimento e considerada como uma função de n, "converge" em certo sentido probabilístico para $P(A)$, quando $n \to \infty$.

A propriedade (5), ligada ao fato do valor da frequência relativa se estabilizar, ou seja, variar cada vez menos, à medida que o número de repetições aumenta, não é uma conclusão matemática e sim algo que pode ser verificado na prática. Esta característica é também conhecida como regularidade estatística.

Probabilidade:

É uma medida numérica que nos permite quantificar a chance de ocorrência de eventos específicos. É expressa como um número entre 0 e 1, onde 0 indica impossibilidade e 1 indica certeza. A probabilidade de um evento E, denotada por P(E), é calculada como o número de resultados favoráveis a E dividido pelo número total de resultados possíveis no espaço amostral.

17.2. Axiomas de probabilidade

Os axiomas de probabilidade, ou os axiomas de Kolmogorov, são um conjunto de definições que estabelecem as propriedades básicas das probabilidades. Eles fornecem a base matemática para o estudo da teoria da probabilidade. Aqui está um resumo dos três axiomas principais:

Axioma da Não-Negatividade:

A probabilidade de um evento é sempre um número não negativo. Formalmente, para qualquer evento E, a probabilidade de E é maior ou igual a zero, $P(E) \geq 0$.

Axioma da Probabilidade Total:

A probabilidade do espaço amostral completo é igual a 1. Em outras palavras, a soma das probabilidades de todos os resultados possíveis no espaço amostral é igual a 1, $P(S)=1$, onde S é o espaço amostral.

Axioma da Aditividade:

Se os eventos são mutuamente exclusivos (não podem ocorrer simultaneamente), então a probabilidade da união desses eventos é igual à soma das probabilidades individuais dos eventos. Formalmente, se E e F são eventos mutuamente exclusivos, então a probabilidade da união de E e F é igual à soma das probabilidades de E e F, $P(E \cup F) = P(E) + P(F)$.

18. PROB - Variáveis Aleatórias Unidimensionais

18.1. Noção Geral de Variável Aleatória

O espaço amostral, conjunto dos resultados possíveis associado a um experimento, não precisa ser necessariamente um conjunto composto por números. Por exemplo, ao descrever uma peça manufaturada, podemos empregar os valores "com defeito" e " sem defeito". No entanto, em muitas situações experimentais, estaremos interessados em resultados expressos por números. Mesmo no exemplo das peças, poderemos atribuir 1 às peças perfeitas e 0 às defeituosas.

Em muitas situações experimentais, desejamos atribuir um número real x a todo elemento s do espaço amostral S. Isto é, x = X(s) é o valor de uma função X do espaço amostral, que liga o conjunto dos valores s, a valores x no espaço dos números reais. Com isto em mente, formulamos a seguinte definição.

Definição. Sejam $\mathcal{E}$ um experimento e S um espaço amostral associado ao experimento. Uma *função* X, que associe a cada elemento $s \in S$ um número real, $X(s)$, é denominada *variável aleatória*.

Importante:

- Na construção da matemática associada aos experimentos estatísticos, é importante entendermos bem a função "variável aleatória", que associa um elemento do espaço amostral a um número real (variável aleatória = função). Reforçando a definição, ao realizarmos um experimento E, associado a um espaço amostral S, que dá um resultado s (que pode ser não numérico) pertencente a S, calculamos o número X(s).
- Quando nós falamos em escolher uma pessoa ao acaso, de alguma população designada, e medimos sua altura (em centímetros, por exemplo), poderemos nos referir aos resultados possíveis como uma variável aleatória X. Mas, uma vez escolhida a pessoa, e medido sua altura, obteremos um valor específico de X, que pode ser designado por x.
- Para um dado valor **r** obtido como resultado de um experimento, sempre haverá o interesse de determinar a probabilidade associada a esse valor: P(X = **r**).

- **OBS**: como veremos adiante, a probabilidade de ocorrência de um valor específico da variável aleatória, só é possível para o caso das variáveis aleatórias discretas. Para o caso das variáveis aleatórias contínuas, só é possível calcular probabilidade de intervalos de valores da variável.

18.2. Variáveis Aleatórias Discretas

Definição: Variáveis aleatórias discretas são aquelas que representam resultados numéricos de um experimento aleatório, onde cada um deles tem uma probabilidade associada. Elas podem assumir um número finito de valores ou um número infinito de valores enumeráveis, ou seja, cada valor pode ser associado a um número natural. Em outras palavras, seus resultados são contáveis e não contínuos. Uma variável aleatória discreta está bem definida se pudermos indicar os possíveis valores x1, x2, ..., xn que ela pode assumir e as respectivas probabilidades p(x1), p(x2), ..., p(xn).

Definição. Seja X uma variável aleatória. Se o número de valores possíveis de X (isto é, R_X, o contradomínio) for finito ou infinito numerável, denominaremos X de *variável aleatória discreta*. Isto é, os *valores* possíveis de X, podem ser postos em lista como $x_1, x_2, \ldots, x_n$. No caso finito, a lista acaba, e no caso infinito numerável, a lista continua indefinidamente.

Exemplos:

- Lançamento de um dado: A variável aleatória representa o número que aparece no dado, que pode ser de 1 a 6.
- Número de filhos em uma família: Aqui, a variável aleatória é o número de filhos em uma família, que pode ser 0, 1, 2, 3, e assim por diante.
- Contagem de carros em um estacionamento: A variável aleatória é o número de carros em um determinado momento, que pode ser qualquer número inteiro não negativo.
- Pontuação em um exame de múltipla escolha: A variável aleatória é o número de perguntas respondidas corretamente, que pode variar de 0 a um número máximo de perguntas no exame.

Características:

- Valores Discretos: Os valores que a variável pode assumir são distintos e separados.
- Função de Probabilidade: Para cada valor possível, existe uma probabilidade associada.
- Soma das Probabilidades: A soma de todas as probabilidades para todos os valores possíveis é igual a 1.
- Exclusão Mútua: A ocorrência de um valor exclui a ocorrência de outros valores possíveis.

Probabilidade associada à ocorrência de valores específicos da variável aleatória.

Definição. Seja X uma variável aleatória discreta. Portando, R_x, o contradomínio de X, será formado no máximo por um número infinito numerável de valores $x_1, x_2, \ldots$ A cada possível resultado x_i associaremos um número $p(x_i) = P(X = x_i)$, denominado probabilidade de x_i. Os números $p(x_i)$, $i = 1, 2, \ldots$ devem satisfazer às seguintes condições:

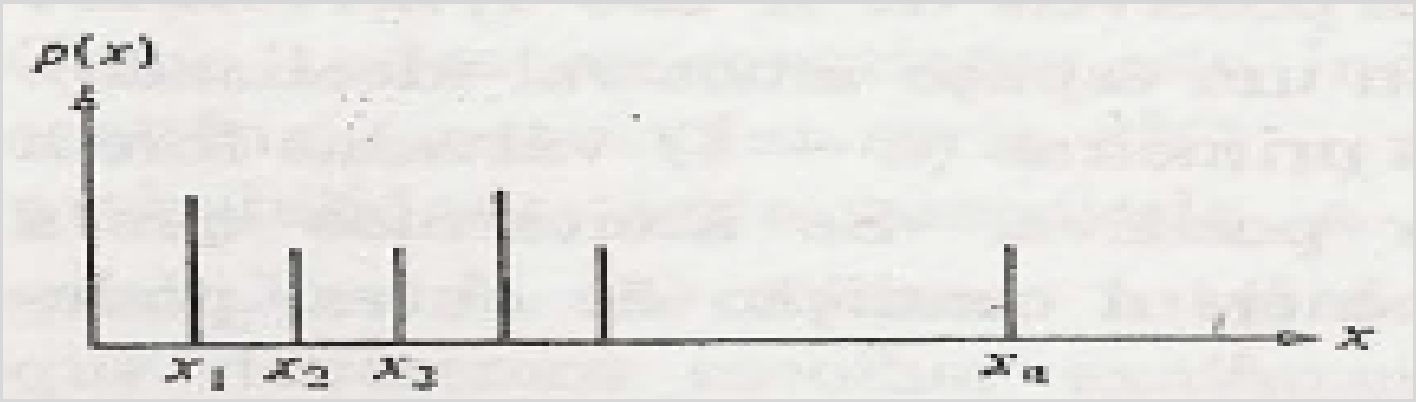
$$(a) \quad p(x_i) \geq 0 \text{ para todo } i,$$

$$(b) \quad \sum_{i=1}^{\infty} p(x_i) = 1. \tag{4.3}$$

A função p, definida acima, é denominada *função de probabilidade* (ou função de probabilidade no ponto) da variável aleatória X. A coleção de pares $[x_i, p(x_i)]$, $i = 1, 2, \ldots$, é algumas vezes denominada *distribuição de probabilidade* de X.

Gráfico de p(x):

- no eixo x estão os resultados possíveis (x_1, x_2, x_3, . . .x_n) do experimento e
- no eixo y estão os valores de probabilidade associados aos resultados



Probabilidade de evento B, subconjunto de S, composto por vários resultados B={x_1, x_2, x_3, . . . }

Seja B um evento associado à variável aleatória X; isto é, $B \subset R_X$ (Fig. 4.5). Suponha-se, especificamente, que $B = \{x_{i_1}, x_{i_2}, \ldots\}$. Daí,

$$P(B) = P[s \mid X(s) \in B] \text{ (porque esses eventos são equivalentes)}$$

$$= P[s \mid X(s) = x_{i_j}, j = 1, 2, \ldots] = \sum_{j=1}^{\infty} p(x_{i_j}). \qquad (4.4)$$

Explicando: A probabilidade de um evento B é igual à soma das probabilidades dos resultados individuais associados com B.

18.3. Variáveis Aleatórias Continuas

Suponha que, ao invés de um conjunto pequeno de valores discretos, como vimos no exemplo das peças, acima (1, 2, 3, . . .), a variável aleatória possa assumir uma grande gama de valores, dentro de um conjunto finito. Por exemplo, valores entre 0 e 1 da forma: 0; 0,01; 0,02; . . .; 0,98; 0,99; 1,00. A cada um destes valores, podemos associar um número não negativo (probabilidade) $p(x_i) = P(X=x_i)$, i = 1,2, 3 . . ., e a soma de todos eles será 1.

Se fizermos a suposição de que os valores da variável aleatória X pode assumir todos os valores possíveis entre 0 e 1, o que acontecerá às probabilidades nos pontos $p(x_i)$? Como os valores de X não são

enumeráveis, não há como falar do *i-ésimo* valor de X, e, por esta razão, $p(x_i)$ é algo que não faz sentido.

Uma variável aleatória contínua é aquela que pode assumir, dentro de um determinado intervalo, infinitos e incontáveis valores (incluindo frações e números irracionais), independentemente do tamanho do intervalo. Em outras palavras, os valores que uma variável aleatória contínua pode assumir são retirados de um conjunto infinito e não podem ser contados individualmente.

Por exemplo, considere a altura das pessoas. A altura pode assumir qualquer valor dentro de um intervalo contínuo, como de 1,50 metros a 2,00 metros. Não é possível listar todos os possíveis valores de altura individualmente, pois há um número infinito de possíveis valores dentro desse intervalo.

Diferentemente das variáveis aleatórias discretas, para as variáveis aleatórias contínuas vale o seguinte:

as probabilidades, associadas a valores específicos delas, valem zero. Mas isso não implica na impossibilidade de ocorrência de valores específicos da variável. Lembrar que, a probabilidade associada a um evento impossível é zero, mas a probabilidade zero não implica, necessariamente, em um evento impossível.

não existem para elas as "funções de probabilidade". O que existem são as funções de "densidade de probabilidade", FDPs, que descreve a probabilidade da variável aleatória assumir um valor dentro de um intervalo específico de valores. As probabilidades são calculadas pela área embaixo da função FDP, dentro desse intervalo.

Faremos, a seguir uma analogia entre o histograma e a função densidade de probabilidade.

Na curva da função densidade de probabilidade, para qualquer ponto (x, y), temos que x é um valor da variável aleatória e y é o valor da densidade de probabilidade no ponto x e não o valor da probabilidade nele.

Em um histograma, temos uma sequência de vários retângulos colados uns nos outros, montados da seguinte maneira.

No eixo x são colocadas as classes (em número > = 5 e <= 20), cujos tamanhos, representando as bases dos retângulos, foram definidas durante o levantamento estatístico descritivo em pauta.

No eixo y são colocados os valores das alturas dos retângulos. Estes valores representam o número de ocorrências (absoluto ou percentual) por "tamanho da classe" (base do retângulo), ou seja, o número de ocorrências verificado dentro da classe. Então, a razão (número de ocorrências)/(tamanho da classe) é uma medida de densidade, ou seja, as alturas dos retângulos no eixo y do histograma são medidas de densidade, semelhantemente ao que ocorre com os valores do eixo y no caso do gráfico de densidade de probabilidade.

No gráfico do histograma os únicos pares (x, y), que fazem sentido, são os que estão no eixo x delimitando os tamanhos das classes. Os demais pares não possuem significado algum. Nem mesmo os pares cujos valores de x estão presentes na tabela de ocorrências levantada.

Da mesma forma, no gráfico de densidade de probabilidade os únicos pares (x, y), que fazem sentido, são os que estão no eixo x delimitando os intervalos para os quais desejamos calcular a probabilidade que, como já vimos, é a área sob a curva delimitada pelo intervalo.

Desta forma, a primeira parte da analogia está descrita, já que tanto no caso da curva de densidade de probabilidade como no caso do histograma, os valores presentes no eixo y são de densidade.

Agora, como chegar aos valores de probabilidade?

No caso da função de densidade de probabilidade, as probabilidades são dadas pelas áreas sob a curva do gráfico, nos intervalos, ou trecho, entre dois valores das variáveis aleatórias (para calcular a área, é necessário calcular a integral da função no trecho).

No caso do histograma, a área do retângulo, que é a altura (número de ocorrências/tamanho da classe) multiplicada pela base (tamanho da classe), resulta no número de ocorrências no intervalo. Agora, se, dentro do intervalo, no lugar de números de ocorrências, consideramos a frequência relativa de ocorrências, (número de ocorrências)/(número total), o valor fracionário resultante para a área será a probabilidade da ocorrência de eventos dentro do intervalo.

Então, como tanto no caso do histograma como no caso da função densidade de probabilidade, as probabilidades são dadas por valores de áreas, a segunda parte da analogia está descrita.

E como é possível chegar à função densidade de probabilidade partindo-se de um histograma?

A explicação vem a seguir, mas antes disso, vamos relembrar uma outra forma de representar as distribuições de frequências, que é o "polígono

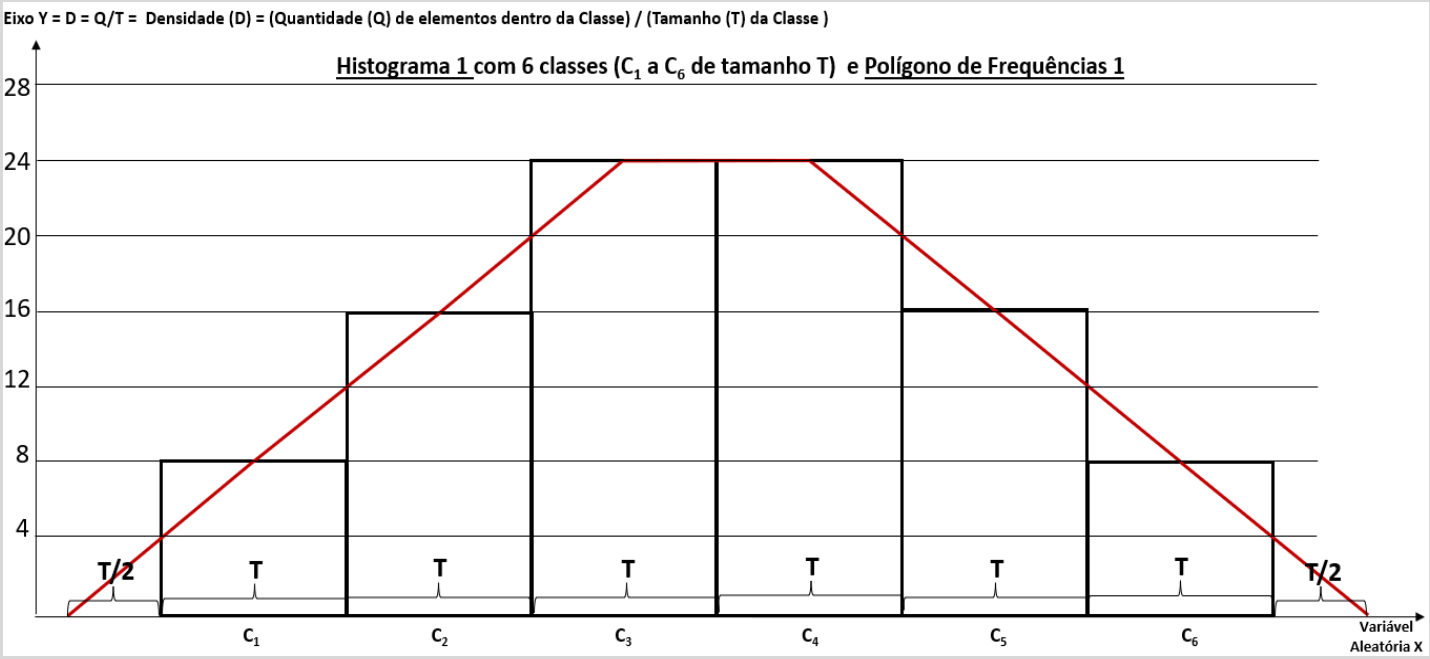
de frequências". Ele é traçado unindo-se os pontos médios dos lados superiores dos retângulos do histograma, com a seguinte particularidade. Os pontos mais à esquerda e mais à direita, localizados no eixo y, são definidos considerando-se os pontos médios de duas classes adicionais, imaginárias, de altura igual a zero, sendo uma delas localizada à esquerda da primeira classe definida para o histograma e a outra localizada depois da última classe.

Se olharmos para o Histograma e para o polígono de frequências, não é difícil perceber que, embora seus contornos sejam diferentes, as suas áreas são iguais (100% ou 1) (FIGURA xxx). Também, é possível observar que, à medida que o número de classes do histograma aumenta, os contornos vão ficando cada vez mais parecidos (ver os gráficos ilustrativos).

Se o comprimento de classe δ, for sendo reduzido até a situação limite em que $\delta \rightarrow 0$, o formato do histograma fica igual ao do polígono de frequências, e nesta condição, a curva resultante é curva de densidade de probabilidade, para a variável aleatória contínua, em questão. A área total sob a curva permanece igual a 1. E a maneira de calcular a probabilidade entre dois valores a e b, como já vimos, é área sob a curva, delimitada pelos pontos a e b.

Vamos, agora, verificar alguns histogramas que ilustram o que foi mencionada acima.

Vamos começar com um histograma de 6 classes.

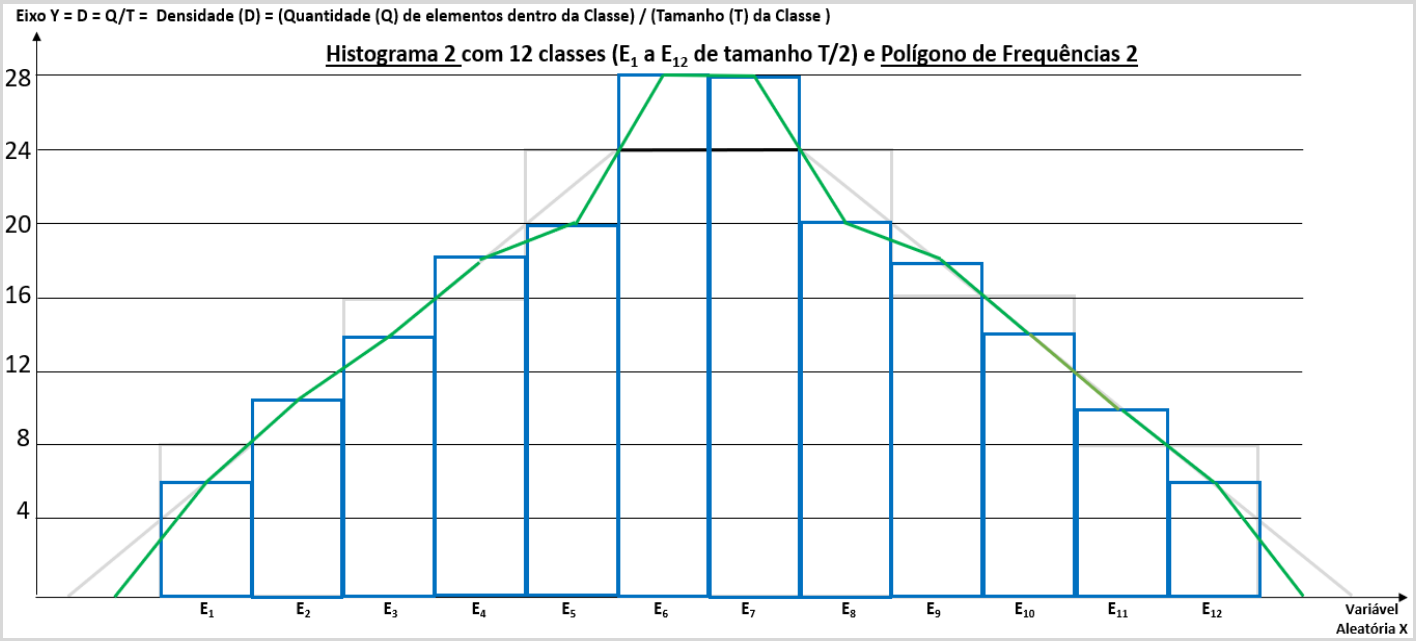


A Figura acima mostra um histograma e um polígono de frequências.

A tabela abaixo resume as características desses gráficos.

Classes	Tamanho de cada Classe = Base do Retângulo da Classe	Densidade de cada Classe = (Quantidade de ocorrências em cada Classe)/(Tamanho da Classe)	Quantidade de Elementos por Classe = Densidade de cada Classe x Tamanho da Classe (T) = Área do Retângulo de cada Classe = Base (T) x Altura	Quantidade Percentual = (Quantidade na Classe)/ (Quantidade Total= 96)
C1	T	8/T	(8/T)T = 8	8/96 =
C2	T	16/T	(16/T)T = 16	16/96 =
C3	T	24/T	(24/T)T = 24	24/96 =
C4	T	24/T	(24/T)T = 24	24/96 =
C5	T	16/T	(16/T)T = 16	16/96 =
C6	T	8/T	(8/T)T = 8	8/96 =

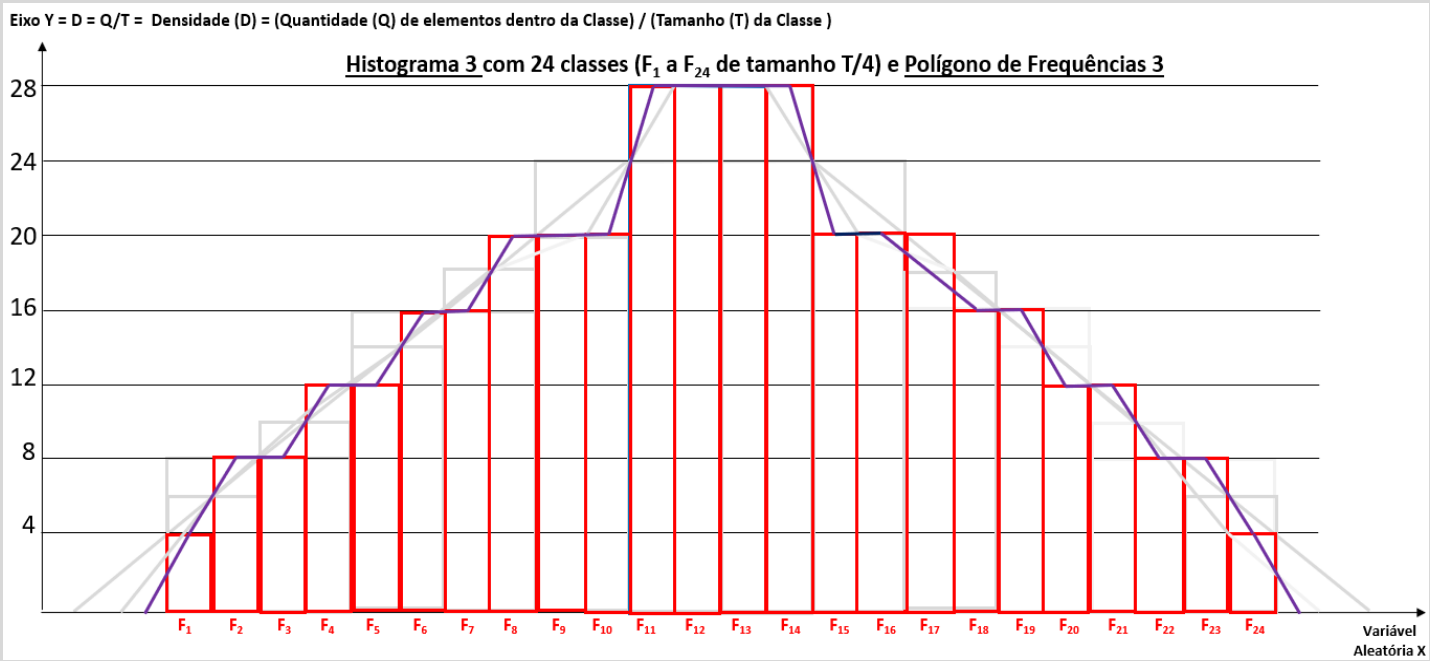
OBS: é importante ressaltar que os valores de densidade (número de elementos da classe/ tamanho T da classe) colocados no eixo y, para o histograma 1, serão mantidos para os histogramas 2 e 3, ou seja, não haverá mudança na escala do eixo y. Então, quando os tamanhos das classes, que no histograma 1 são 6 (C1 a C6) de tamanho T, forem alterados para T/2, no histograma 2 (com 12 classes de E1 a E12) e T/4, no histograma 3 (com 24 classes de F1 a F24), os cálculos das densidade continuarão a ser n/T (cravado no histograma 1), como poderemos ver nas tabelas associadas aos casos do histogramas 1 e 2.



A Figura acima mostra um histograma e um polígono de frequências, para um número de classes igual a 12. Nela, é possível verificar que os formatos das áreas sob o histograma e sob o polígono de frequências são mais próximos entre si, do que os mesmos formatos da Figura anterior, onde o número de classes era igual a 6.

A tabela abaixo resume as características desses gráficos.

Classes	Tamanho de cada Classe = Base do Retângulo da Classe	Densidade de cada Classe = (Quantidade de ocorrências em cada Classe)/ (Tamanho da Classe)	Quantidade de Elementos por Classe = Densidade de cada Classe x Tamanho da Classe (T) = Área do Retângulo de cada Classe = Base (T) x Altura	Quantidade Percentual = (Quantidade na Classe)/ (Quantidade Total= 96)
E1	T/2	6/T	(6/T)X(T/2) = 3	3/96 =
E2	T/2	10/T	(10/T)X(T/2) = 5	5/96 =
E3	T/2	14/T	(14/T)X(T/2) = 7	7/96 =
E4	T/2	18/T	(18/T)X(T/2) = 9	9/96 =
E5	T/2	20/T	(20/T)X(T/2) = 10	10/96 =
E6	T/2	28/T	(28/T)X(T/2) = 14	14/96 =
E7	T/2	28/T	(28/T)X(T/2) = 14	14/96 =
E8	T/2	20/T	(20/T)X(T/2) = 10	10/96 =
E9	T/2	18/T	(18/T)X(T/2) = 9	9/96 =
E10	T/2	14/T	(14/T)X(T/2)= 7	7/96 =
E11	T/2	10/T	(10/T)X(T/2) = 5	5/96 =
E12	T/2	4/T	(6/T)X(T/2) = 3	3/96 =



A Figura acima mostra um histograma e um polígono de frequências, para um número de classes igual a 24. Nela, é possível verificar que os formatos das áreas sob o histograma e sob o polígono de frequências são ainda mais próximos entre si, do que os mesmos formatos da Figura anterior, onde o número de classe era igual a 12.

A tabela abaixo resume as características desses gráficos.

Classes	Tamanho de cada Classe = Base do Retângulo da Classe	Densidade de ocorrências em cada Classe = (Quantidade de cada Classe)/ (Tamanho da Classe)	Quantidade de Elementos por Classe = Densidade de cada Classe x Tamanho da Classe (T) = Área do Retângulo de cada Classe = Base (T) x Altura	Quantidade Percentual = (Quantidade na Classe)/ (Quantidade Total= 96)
F1	T/4	4/T	(4/T)x(T/4) = 1	1/96 =
F2	T/4	4/T	(8/T)x(T/4) = 2	2/96 =
F3	T/4	8/T	(8T)x(T/4) = 2	2/96 =
F4	T/4	12/T	(12/T)x(T/4) = 3	3/96 =
F5	T/4	12/T	(12/T)x(T/4) = 3	3/96 =
F6	T/4	16/T	(16/T)x(T/4) = 4	4/96 =
F7	T/4	18/T	(16/T)x(T/4) = 4	4/96 =
F8	T/4	18/T	(20/T)x(T/4) = 5	5/96 =
F9	T/4	18/T	(20/T)x(T/4) = 5	5/96 =
F10	T/4	22/T	(20/T)x(T/4)= 5	5/96 =
F11	T/4	28/T	(28/T)x(T/4) = 7	7/96 =
F12	T/4	28/T	(28/T)x(T/4) = 7	7/96 =
F13	T/4	28/T	(28/T)x(T/4) = 7	7/96 =
F14	T/4	28/T	(28/T)x(T/4) = 7	7/96 =
F15	T/4	22/T	(20T)x(T/4) = 5	5/96 =
F16	T/4	18/T	(20/T)x(T/4) = 5	5/96 =
F17	T/4	18/T	(16/T)x(T/4) = 4	4/96 =
F18	T/4	18/T	(16/T)x(T/4) = 4	4/96 =
F19	T/4	16/T	(12/T)x(T/4) = 3	3/96 =
F20	T/4	12/T	(12/T)x(T/4) = 3	3/96 =
F21	T/4	12/T	(12/T)x(T/4) = 3	3/96 =

F22	T/4	8/T	(8/T)x(T/4)= 2	2/96 =
F23	T/4	4/T	(8/T)x(T/4) = 2	2/96 =
F24	T/4	4/T	(4/T)x(T/4) = 1	1/96 =

Então, através dos 3 histogramas e polígonos de frequência mostrados anteriormente, é possível constatar a tendência que existe do histograma e do polígono de frequências se converterem na curva de densidade de probabilidade, se o comprimento de classe δ, for sendo reduzido até a situação limite em que $\delta \rightarrow 0$.

18.4. Variáveis Aleatórias Discretas Importantes (revisar!)

- Variável Aleatória Bernoulli: Uma variável aleatória que assume apenas dois valores possíveis, geralmente denotados como 0 e 1. É frequentemente usada para modelar eventos binários, como sucesso ou falha em experimentos.

- Variável Aleatória Binomial: É o número de sucessos em um número fixo de tentativas independentes, onde cada tentativa tem a mesma probabilidade de sucesso. É uma generalização da variável aleatória Bernoulli e é usada para modelar situações em que ocorrem múltiplos eventos de Bernoulli.

- Variável Aleatória de Poisson: Modela o número de eventos que ocorrem em um intervalo de tempo fixo ou em uma região fixa, se os eventos ocorrerem a uma taxa média conhecida e independentemente um do outro. É frequentemente usado em problemas de contagem, como modelar o número de chamadas recebidas por uma central telefônica em um determinado período de tempo.

- Variável Aleatória Geométrica: Representa o número de tentativas independentes necessárias para obter o primeiro sucesso em um processo de Bernoulli, onde cada tentativa tem a mesma probabilidade de sucesso.

- Variável Aleatória de Distribuição Hipergeométrica: Modela o número de sucessos em uma amostra sem reposição, onde a população tem um número finito de elementos divididos em duas categorias distintas (sucesso e falha).

- Variável Aleatória de Uniforme Discreta: Assume um número finito de valores igualmente espaçados dentro de um intervalo específico com a mesma probabilidade de ocorrência para cada valor.

- Variável Aleatória de Distribuição de Probabilidade Multinomial: É uma generalização da distribuição binomial para mais de duas categorias. É usada quando há mais de duas categorias possíveis e cada tentativa resulta em uma dessas categorias.

18.5. Variáveis Aleatórias Contínuas Importantes (revisar!)

- Variável Aleatória Uniforme Contínua: Assume valores dentro de um intervalo específico com uma distribuição uniforme, onde cada valor dentro do intervalo tem a mesma probabilidade de ocorrência.

- Variável Aleatória Normal (Gaussiana): É uma das distribuições mais importantes na teoria das probabilidades e estatística. A distribuição normal é simétrica em torno de sua média e tem uma forma de sino. É amplamente utilizada devido ao seu papel central no teorema central do limite e em muitos modelos estatísticos.

- Variável Aleatória Exponencial: Modela o tempo entre eventos em um processo de Poisson, onde os eventos ocorrem continuamente e independentemente uns dos outros. A distribuição exponencial é frequentemente usada em teoria de filas, tempo de vida de produtos e em análises de confiabilidade.

- Variável Aleatória de Distribuição de Cauchy: É uma distribuição simétrica que tem uma forma similar à distribuição normal, mas com caudas mais longas. É usada em estatísticas bayesianas e em alguns contextos da física.

- Variável Aleatória de Distribuição de Pareto: É uma distribuição de caudas pesadas que é comumente usada em modelos que envolvem fenômenos onde a maioria dos eventos tem valores pequenos, mas alguns eventos têm valores muito grandes.

- Variável Aleatória de Distribuição de Weibull: É usada para modelar o tempo de vida de objetos ou sistemas. Pode descrever sistemas que falham com taxas constantes ou crescentes ou decrescentes.

- Variável Aleatória de Distribuição Log-Normal: É usada para modelar variáveis que são o resultado do processo de exponenciação de uma variável normalmente distribuída. É comumente usada para modelar preços de ativos financeiros, tamanhos de partículas e outras grandezas positivas que não podem assumir valores negativos.

18.6. Funções de Variável Aleatória (revisar!)

As funções de variáveis aleatórias são transformações aplicadas a variáveis aleatórias que geram novas variáveis aleatórias. Essas funções são fundamentais em estatística e probabilidade, pois permitem modelar e analisar diferentes aspectos dos dados. Aqui estão algumas das funções de variáveis aleatórias mais comuns:

- Função de Distribuição Cumulativa (CDF): Para qualquer variável aleatória X, a função de distribuição cumulativa F(x) é definida como a probabilidade de que X assuma um valor menor ou igual a x. Ela fornece uma descrição completa da distribuição de probabilidade de X.

- Função de Densidade de Probabilidade (PDF): Para variáveis aleatórias contínuas, a função de densidade de probabilidade f(x) descreve a densidade de probabilidade em torno de um determinado valor x. A área sob a curva da PDF em um intervalo dá a probabilidade de que a variável aleatória caia dentro desse intervalo.

- Função de Massa de Probabilidade (PMF): Para variáveis aleatórias discretas, a função de massa de probabilidade p(x) fornece a

probabilidade de que a variável aleatória assuma um valor específico x.

- Função de Sobrevivência (Survival Function): A função de sobrevivência S(x) é complementar à CDF e é definida como a probabilidade de que a variável aleatória exceda um determinado valor x.

- Função de Quantil (Quantile Function): A função de quantil Q(p) é o valor que divide a distribuição de probabilidade em uma proporção p. Por exemplo, o quantil de ordem 0.5 é a mediana da distribuição.

- Função Geradora de Momentos (Moment Generating Function, MGF): É uma função que permite calcular momentos de uma variável aleatória. Ela fornece uma maneira eficiente de derivar momentos de ordem superior.

- Função Característica: Similar à MGF, a função característica é uma transformada de Fourier da PDF da variável aleatória. Ela fornece informações sobre todos os momentos da distribuição.

18.7. Variáveis Aleatórias de Duas ou Mais Dimensões (revisar!)

Variáveis aleatórias de duas ou mais dimensões referem-se a conjuntos de variáveis aleatórias que estão relacionadas de alguma forma. Elas são fundamentais em estatística e teoria das probabilidades para modelar situações mais complexas onde múltiplas variáveis estão envolvidas. Aqui estão algumas das principais classes de variáveis aleatórias de duas ou mais dimensões:

- Variáveis Aleatórias Conjuntas: Este é o caso mais simples de variáveis aleatórias de duas dimensões. Uma variável aleatória conjunta consiste em duas ou mais variáveis aleatórias que estão associadas em uma distribuição conjunta. Por exemplo, se tivermos duas variáveis aleatórias X e Y, a distribuição conjunta descreveria a relação entre X e Y.

- Variáveis Aleatórias Condicionais: São variáveis aleatórias que estão condicionadas a certos valores de outras variáveis. Por exemplo, se tivermos duas variáveis aleatórias X e Y, a variável aleatória condicional X|Y representa a distribuição de X dado que Y assumiu um valor específico.

- Variáveis Aleatórias Independentes: Duas variáveis aleatórias são independentes se a ocorrência de eventos em uma não influenciar a ocorrência de eventos na outra. Isso se estende para variáveis aleatórias de múltiplas dimensões. Se X e Y são independentes, então a distribuição conjunta de X e Y é o produto das suas distribuições marginais.

- Variáveis Aleatórias Multivariadas: São conjuntos de variáveis aleatórias que são tratadas como um único objeto. Por exemplo, um vetor de variáveis aleatórias X = (X1, X2, ..., Xn) é uma variável aleatória multivariada. A distribuição multivariada descreve a relação conjunta entre todas as variáveis nesse vetor.

- Covariância e Correlação: São medidas de associação entre duas variáveis aleatórias. A covariância mede o grau de variabilidade

conjunta entre duas variáveis, enquanto a correlação é a covariância normalizada, fornecendo uma medida padronizada da relação linear entre as variáveis.

- Função de Distribuição Conjunta: Assim como em variáveis univariadas, a função de distribuição conjunta fornece a probabilidade de que as variáveis aleatórias de duas ou mais dimensões assumam determinados valores ou intervalos de valores simultaneamente.

18.8. Caracterização Adicional das Variáveis Aleatórias (revisar!)

Além das descrições e funções mencionadas anteriormente, as variáveis aleatórias podem ser caracterizadas de várias outras formas e cada uma delas fornece informações úteis sobre a sua distribuição e comportamento. Aqui estão algumas considerações adicionais:

- Momentos: Os momentos de uma variável aleatória são medidas que resumem sua distribuição. O momento de ordem k de uma variável aleatória X é dado por E[X^k], onde E[] representa o operador de esperança matemática. O primeiro momento é a média (ou esperança), o segundo momento é a variância, e assim por diante.

- Assimetria e Curtose: A assimetria (skewness) mede a falta de simetria da distribuição de uma variável aleatória. Uma distribuição simétrica tem assimetria zero. A curtose (kurtosis) mede o grau de "picosidade" ou "achatamento" da distribuição em comparação com uma distribuição normal.

- Entropia: A entropia é uma medida de incerteza associada a uma variável aleatória. Quanto maior a entropia, maior é a incerteza. A entropia é frequentemente usada em teoria da informação e em aplicações de codificação de dados.

- Momento Característico e Função Geradora de Momentos (MGF): O momento característico de uma variável aleatória é uma função

relacionada com a transformada de Fourier da função de densidade de probabilidade. Ele fornece uma maneira eficiente de calcular os momentos de ordem superior. A função geradora de momentos (MGF) é uma forma de caracterizar completamente a distribuição de uma variável aleatória através de uma função.

- Função Característica: A função característica é outra maneira de caracterizar completamente a distribuição de uma variável aleatória. É a transformada de Fourier da função de densidade de probabilidade e fornece informações sobre todos os momentos da distribuição.

- Função de Autocorrelação: Em séries temporais ou processos estocásticos, a função de autocorrelação mede a relação entre os valores da variável aleatória em diferentes momentos no tempo. É uma medida importante para entender a dependência temporal em dados.

18.9. Somas de Variáveis Aleatórias (revisar!)

As somas de variáveis aleatórias são de grande importância em estatística e probabilidade, especialmente em contextos onde estamos interessados no resultado combinado de múltiplas variáveis aleatórias. Aqui estão algumas considerações importantes sobre somas de variáveis aleatórias:

- Soma de Variáveis Aleatórias Independentes: Se X e Y são variáveis aleatórias independentes, a distribuição da soma Z = X + Y é dada pela convolução das distribuições de X e Y. Isso significa que para calcular a distribuição de Z, você precisa convolver as funções de densidade de probabilidade de X e Y.

- Média e Variância da Soma: Se X e Y são variáveis aleatórias com médias μX e μY e variâncias σX^2 e σY^2, respectivamente, então a média da soma Z = X + Y é μZ = μX + μY e a variância da soma é σZ^2 = σX^2 + σY^2. Isso vale apenas para variáveis aleatórias independentes.

- Teorema Central do Limite: O teorema central do limite afirma que, para uma grande amostra de variáveis aleatórias independentes e idênticamente distribuídas (iid), a distribuição da soma tende a se aproximar de uma distribuição normal, independentemente da distribuição original das variáveis.

- Soma de Variáveis Aleatórias Correlacionadas: Se as variáveis aleatórias não são independentes, a distribuição da soma pode ser mais complexa e pode depender da natureza da correlação entre elas. Nesses casos, as propriedades da soma podem ser determinadas usando técnicas específicas, como a transformação de Laplace ou a função geradora de momentos.

- Somas de Variáveis Aleatórias em Séries Temporais: Em séries temporais, somas de variáveis aleatórias são comuns ao modelar o comportamento acumulado ao longo do tempo, como somas de retornos financeiros ao longo de vários períodos.

- Somas de Variáveis Aleatórias em Processos de Contagem: Em processos de contagem, como o processo de Poisson, as somas de variáveis aleatórias representam o número total de ocorrências de eventos em um intervalo de tempo ou em uma região específica.

18.10. Probabilidade condicional e independência

A probabilidade condicional nos permite avaliar a probabilidade de um evento, considerando que outro evento já ocorreu. Vamos considerar dois eventos, um evento A e um outro B, sendo que B ocorre primeiro que o A. Nesta condição, a ocorrência do evento B acaba limitando as possibilidades completas do evento A. Ou seja, neste caso só faz sentido olhar as possibilidades de A dentro de B. Assim, a probabilidade condicional, probabilidade de A dado o evento B, denotada por P(A|B), é P(A|B) = P(A inter B)/P(B). Aqui a ocorrência de A é dependente da ocorrência de B.

A independência entre eventos significa que um evento não influencia a ocorrência de outro.

Probabilidade Condicional

A probabilidade condicional de um evento A dado que outro evento B ocorreu é a probabilidade de A ocorrer assumindo que B já ocorreu. Isso é denotado por $P(A|B)$ e é calculado pela fórmula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

desde que $P(B) > 0$.

- Aqui, $P(A \cap B)$ é a probabilidade de ambos os eventos A e B ocorrerem (a interseção de A e B), e $P(B)$ é a probabilidade de B ocorrer.
- A probabilidade condicional reflete como a ocorrência de um evento afeta a probabilidade de ocorrência de outro evento.

Independência

Dois eventos A e B são independentes se a ocorrência de um não afeta a probabilidade de ocorrência do outro. Matematicamente, A e B são independentes se e somente se:

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

Uma consequência importante da independência é que, se A e B são independentes, então:

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

- Isso significa que a probabilidade de A ocorrer, dado que B ocorreu, é simplesmente a probabilidade de A ocorrer, e vice-versa, o que reflete a ideia de que a ocorrência de um evento não afeta a probabilidade do outro.

Exemplos Práticos

Probabilidade Condicional: Se você tem um baralho de 52 cartas e quer saber a probabilidade de puxar um ás dado que a carta que você puxou é de espadas, você está lidando com probabilidade condicional.

Independência: Jogar um dado e girar uma roleta são exemplos de eventos independentes, pois o resultado de um não afeta o resultado do outro.

18.11. Teorema de Bayes

O teorema de Bayes é um conceito fundamental em estatística e probabilidade, usado para atualizar nossas crenças sobre um evento após observarmos evidências relevantes. Ele é formulado matematicamente da seguinte maneira:

Seja A e B dois eventos, com P(B) > 0 (probabilidade de B ser não nula). Então, o teorema de Bayes afirma que a probabilidade condicional de A dado B (ou seja, a probabilidade de que o evento A ocorra, dado que o evento B ocorreu) é dada por:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Onde:

- $P(A|B)$ é a probabilidade condicional de A dado B.
- $P(B|A)$ é a probabilidade condicional de B dado A.
- $P(A)$ e $P(B)$ são as probabilidades marginais de A e B, respectivamente.

O teorema de Bayes é útil quando queremos atualizar nossas crenças sobre a ocorrência de um evento (A) com base em novas evidências (B). Ele fornece uma maneira de relacionar a probabilidade de A dado B com a probabilidade de B dado A, permitindo-nos fazer inferências probabilísticas mais precisas.

O teorema de Bayes é amplamente utilizado em várias áreas, incluindo aprendizado de máquina, diagnóstico médico, reconhecimento de padrões, entre outros. Ele é especialmente útil em situações onde temos dados observados e queremos inferir sobre os parâmetros desconhecidos de um modelo probabilístico.

19. PROB - Distribuições de Probabilidades Discretas

19.1. Distribuições Discretas – Distribuição BINOMIAL

A distribuição binomial é comumente aplicada para modelar situações de experimentos aleatórios que resultam em apenas dois resultados possíveis, como sucesso ou fracasso, sim ou não, cabeça ou coroa.

Exemplo 4.7. Suponha que peças saiam de uma linha de produção e sejam classificadas como defeituosas (D) ou como não-defeituosas (N), isto é, perfeitas. Admita que três dessas peças, da produção de um dia, sejam escolhidas ao acaso e classificadas de acordo com esse esquema. O espaço amostral para esse experimento, S, pode ser assim, apresentado:

$$S = \{DDD, DDN, DND, NDD, NND, NDN, DNN, NNN\}.$$

Suponhamos que seja 0,2 a probabilidade de uma peça ser defeituosa e 0,8 a de ser não-defeituosa.

Peça 1	Peça 2	Peça 3	Número de Peças c/ Defeito	Probabilidade
D = 0,2	D = 0,2	D = 0,2	3	p(3) =0,2x02x0,2 = $(0,2)^3$
D = 0,2	D = 0,2	N=0,8	2	p(2) = 3x0,2x0,8x0,2 = $3x(0,2)^2.(0,8)$
D = 0,2	N=0,8	D = 0,2		
N=0,8	D = 0,2	D = 0,2		
N=0,8	N=0,8	D = 0,2	1	p(1) =3x 0,8x0,2x0,8 = $3x(0,8)^2.(0,2)$
N=0,8	D = 0,2	N=0,8		
D = 0,2	N=0,8	N=0,8		
N=0,8	N=0,8	N=0,8	0	p(0) =0,8x0,8x0,8 = $(0,8)^3$

A soma das probabilidades é p(0) + p(1) + p(2) + p(3) = $(0,8)^3$ + $3x(0,8)^2.(0,2)$ + $3x(0,2)^2.(0,8)$ + $(0,2)^3$ = $(0,8 + 0,2)^3$ = 1

Resumo sobre o caso acima:

- o experimento é analisar a sanidade (perfeição ou defeito) de 3 peças produzidas em uma manufatura;
- o espaço amostral é S = {DDD, DDN, DND, NDD, NND, NDN, DNN, NNN};
- em termos de número de defeitos de placas, o espaço amostral é {0, 1, 2, 3};

- as probabilidades associadas são: p(0) = $(0,8)^3$, p(1) =3x$(0,8)^2$.(0,2), p(2) = 3x$(0,2)^2$.(0,8), p(3) $(0,2)^3$.
- a soma das probabilidades é 1.
- e forma equivalente, esse caso poderia ser caracterizado como uma repetição tripla de:
    - um experimento único (de verificar a sanidade de uma placa);
    - com um espaço amostral S1 = {N, D},
    - com p(N) = 0,8 e p(D) = 0,2

Generalizando:

Definição: Consideremos um experimento $\mathcal{E}$ e seja A algum evento associado a $\mathcal{E}$. Admita-se que $P(A) = p$ e conseqüentemente $P(\overline{A}) = 1 - p$. Considerem-se n repetições de $\mathcal{E}$. Daí, o espaço amostral será formado por todas as seqüências possíveis $\{a_1, a_2, \ldots, a_n\}$, onde cada a_i é ou A ou $\overline{A}$, dependendo de que tenha ocorrido A ou $\overline{A}$ na i-ésima repetição de $\mathcal{E}$. (Existem 2^n dessas seqüências.) Além disso, suponha-se que $P(A) = p$ permaneça a mesma para todas as repetições. A variável aleatória X será assim definida: X = número de vezes que o evento A tenha ocorrido. Denominaremos X de variável aleatória *binomial*, com parâmetros n e p. Seus valores possíveis são evidentemente 0, 1, 2, . . ., n. (De maneira equivalente, diremos que X tem uma *distribuição binomial*.)

As repetições individuais de $\mathcal{E}$ serão denominadas *Provas de Bernouilli*.

Teorema:

Teorema 4.1. Seja X uma variável binomial, baseada em n repetições. Então,

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad k = 0, 1, \ldots, n. \qquad (4.5)$$

Demonstração: Considere-se um particular elemento do espaço amostral de $\mathcal{E}$ satisfazendo à condição $X = k$. Um resultado como esse poderia surgir, por exemplo, se nas primeiras k repetições de $\mathcal{E}$ ocorresse A, enquanto nas últimas $n - k$ repetições ocorresse $\overline{A}$, isto é,

$$\underbrace{AAA \cdots A}_{k} \underbrace{\overline{A}\overline{A}\overline{A} \cdots \overline{A}}_{n-k}.$$

Como todas as repetições são independentes, a probabilidade desta seqüência particular seria $p^k(1 - p)^{n-k}$, mas exatamente essa mesma probabilidade seria associada a qualquer outro resultado para o qual $X = k$. O número total de tais resultados é igual a $\binom{n}{k}$, porque deveremos escolher exatamente k posições (dentre n) para o evento A. Ora, isso dá o resultado acima, porque esses $\binom{n}{k}$ resultados são todos mutuamente excludentes.

3. **Distribuição Binomial:**
   - A distribuição binomial modela o número de sucessos em um número fixo de tentativas independentes. Sua função de densidade de probabilidade é:

     $f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$

     Onde:

     n - número total de tentativas,

     p - probabilidade de sucesso em uma tentativa.

Seguem alguns exemplos de fenômenos que seguem a distribuição binomial:

- Lançamento de uma Moeda: Cada lançamento de uma moeda pode resultar em sucesso (cara) ou fracasso (coroa). Se estivermos interessados em contar o número de caras em uma série de lançamentos, podemos modelar isso usando uma distribuição binomial.

- Teste de Sucesso/Fracasso: Em um teste de múltipla escolha onde cada questão tem apenas duas opções de resposta (certo ou errado), podemos modelar o número de respostas corretas de um aluno como uma distribuição binomial.

- Taxa de Resposta em Campanhas de Marketing: Em uma campanha de marketing direto, onde os clientes podem responder (sucesso) ou não responder (fracasso), podemos modelar a taxa de resposta como uma distribuição binomial.

- Produção Industrial de Itens Defeituosos: Em uma linha de produção, cada item produzido pode ser classificado como defeituoso ou não defeituoso. Se estivermos interessados em calcular a probabilidade de um certo número de itens defeituosos em uma amostra, podemos usar uma distribuição binomial.

- Resultado de uma Eleição Binária: Em uma eleição onde há apenas dois candidatos, podemos modelar o número de votos para um candidato como uma distribuição binomial.

19.2. Distribuições discretas – Distribuição de Bernoulli

Uma V.A. (X) de Bernoulli é aquela que assume apenas dois valores **1** se ocorrer **sucesso** (S) e **0** se ocorrer **fracasso** (F), com probabilidade de sucesso **p**, isto é,

$$X = \begin{cases} 1, \text{ se ocorrer "sucesso"} \\ 0, \text{ se ocorrer "fracasso"} \end{cases}$$

E sua função de probabilidade é dada por:

x	0	1
P(X=x)	1-p	p

Notação: X~Bernoulli (p), indica que a v.a. X tem distribuição de Bernoulli com parâmetro **p**

Se X~Bernoulli(p) pode-se mostrar que: E(X)=p e Var(X)=p(1-p).

A distribuição de Bernoulli é uma distribuição de probabilidade discreta que modela experimentos aleatórios com apenas dois resultados possíveis: sucesso ou fracasso, onde o sucesso tem probabilidade p e o fracasso tem probabilidade 1-p.

Aqui estão alguns exemplos de fenômenos que podem ser modelados pela distribuição de Bernoulli:

- Lançamento de uma moeda justa: Se considerarmos "cara" como sucesso e "coroa" como fracasso, então o resultado de um único lançamento de uma moeda justa pode ser modelado usando a distribuição de Bernoulli.
- Teste de sucesso ou falha: Se um dispositivo é testado para determinar se funciona corretamente ou não, o resultado do teste pode ser modelado como uma distribuição de Bernoulli, onde "sucesso" indica que o dispositivo está funcionando corretamente e "fracasso" indica uma falha.
- Resposta a uma pergunta de sim ou não: Em pesquisas ou questionários, quando uma pergunta é formulada de forma que a resposta seja apenas "sim" ou "não", a distribuição de respostas pode ser modelada pela distribuição de Bernoulli.
- Resultados de um experimento médico binário: Em um experimento clínico onde um tratamento é administrado e o resultado é "cura" ou "não cura", a distribuição de Bernoulli pode ser usada para modelar a probabilidade de cura para cada paciente.

- Resultado de um evento esportivo de duas equipes: Em esportes como futebol, basquete ou tênis, onde uma equipe ou jogador pode vencer ou perder, os resultados de jogos individuais podem ser modelados usando a distribuição de Bernoulli.

Esses são alguns exemplos de situações onde a distribuição de Bernoulli é aplicável, representando experimentos com apenas dois resultados possíveis: sucesso ou fracasso.

19.3. Distribuições discretas – Distribuição de Poisson

Distribuição de Poisson: é uma distribuição de probabilidade discreta que expressa a probabilidade de um determinado número de eventos (1, 2, 3, 4 . . .) ocorrer em um intervalo fixo de tempo ou espaço, sendo que:

- estes eventos são independentes, ou seja, o tempo de ocorrência de qualquer um deles independe do tempo de ocorrência de um evento anterior;
- a taxa média de ocorrências (**a**) é conhecida.

Dadas as condições acima, a função de distribuição de probabilidade é dada pela equação abaixo:

$$P(X = k) = \frac{e^{-a}\, a^k}{k!}, \qquad k = 0, 1, \ldots, n, \ldots,$$

Onde:

a = é a taxa média de ocorrência do evento (**número de ocorrência / tempo fixo**);

k = quantidade de eventos em um tempo fixo;

P(X = k) = probabilidade de ocorrência de **k** eventos em um intervalo de **tempo fixo**

Exemplos de fenômenos que seguem a distribuição de Poisson:

- Número de chamadas recebidas em um call center: Se a taxa média de chamadas por hora é conhecida, o número de chamadas que chegam em um determinado intervalo de tempo pode ser modelado pela distribuição de Poisson.

- Número de acidentes de trânsito em uma interseção: Se sabemos a taxa média de acidentes por dia em uma interseção, podemos usar a distribuição de Poisson para modelar o número de acidentes que ocorrem em um determinado período de tempo.
- Número de defeitos em um lote de produtos: Se a taxa média de defeitos por unidade é conhecida, podemos usar a distribuição de Poisson para modelar o número de defeitos em um lote de produtos.
- Número de partículas emitidas por uma fonte radioativa em um intervalo de tempo: Se a taxa média de emissão de partículas por segundo é conhecida, o número de partículas emitidas em um determinado período de tempo pode ser modelado pela distribuição de Poisson.
- Número de eventos sísmicos em uma região: Se a taxa média de terremotos por ano é conhecida para uma determinada região, podemos usar a distribuição de Poisson para modelar o número de terremotos que ocorrem em um período de tempo específico.

No gráfico da função de distribuição de probabilidade, no eixo x são colocados os valores de da variável aleatória, quantidades de eventos (1, 2, 3, 4, . . .). As variáveis podem ser algumas das várias mencionadas no exemplo acima (número de ocorrências em um intervalo de tempo, espaço ou volume, com uma taxa média conhecida **a**, conhecida). As probabilidades, P(X = 1), P(X = 2), associadas às variáveis aleatórias são colocadas no eixo y.

Para esta, e para outras distribuições discretas, as várias probabilidades são definidas em pontos específicos (valores no eixo x). Diferentemente do que ocorre com as distribuições associadas às variáveis aleatórias contínuas. Nestas, as probabilidades são determinadas em trechos definidos por dois valores da v.a., através do cálculo da área sob a curva, da função densidade de probabilidade, compreendida no trecho.

4. **Distribuição de Poisson:**
   - A distribuição de Poisson modela o número de eventos raros em um intervalo de tempo ou espaço fixo. Sua função de densidade de probabilidade é:

     $f(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$

     Onde:

     λ - taxa média de ocorrência dos eventos.

Exemplos de fenômenos que seguem a distribuição de Poisson:

A distribuição de Poisson é comumente encontrada em situações onde estamos interessados no número de eventos que ocorrem em um intervalo fixo de tempo, espaço ou volume, quando esses eventos ocorrem de forma independente e a uma taxa média conhecida. Aqui estão alguns exemplos de fenômenos que seguem a distribuição de Poisson:

- Número de chamadas recebidas em um call center: Se a taxa média de chamadas por hora é conhecida, o número de chamadas que chegam em um determinado intervalo de tempo pode ser modelado pela distribuição de Poisson.
- Número de acidentes de trânsito em uma interseção: Se sabemos a taxa média de acidentes por dia em uma interseção, podemos usar a distribuição de Poisson para modelar o número de acidentes que ocorrem em um determinado período de tempo.
- Número de defeitos em um lote de produtos: Se a taxa média de defeitos por unidade é conhecida, podemos usar a distribuição de Poisson para modelar o número de defeitos em um lote de produtos.
- Número de partículas emitidas por uma fonte radioativa em um intervalo de tempo: Se a taxa média de emissão de partículas por segundo é conhecida, o número de partículas emitidas em um determinado período de tempo pode ser modelado pela distribuição de Poisson.
- Número de eventos sísmicos em uma região: Se a taxa média de terremotos por ano é conhecida para uma determinada região, podemos usar a distribuição de Poisson para modelar o número de terremotos que ocorrem em um período de tempo específico.

No gráfico da função da distribuição de probabilidade de Poisson, no eixo x são colocados os valores da variável aleatória (v.a.) de interesse, que pode ser uma das várias mencionadas dentre os exemplos acima (número de ocorrências em um intervalo de tempo, espaço ou volume, com uma taxa média conhecida **l**, conhecida). Pelo fato da distribuição de Poisson estar associada a variáveis aleatórias discretas, o gráfico da função de distribuição de probabilidade não é contínuo, sendo que no eixo x estão colocadas as varáveis aleatórias e no eixo y estão as probabilidades associadas. Ou seja, neste caso não faz sentido o uso do termo "função densidade de probabilidade" que, como veremos mais à frente, deve ser

usado no caso das variáveis aleatórias contínuas. Porque, para estas, a única probabilidade capaz de ser calculada é a da variável aleatória estar contida dentro de um intervalo definido no eixo x. Neste caso, a probabilidade é dada pela área sob a curva, delimitada pelo intervalo. Em outras palavras, o cálculo é feito através da integral da função densidade de probabilidade no intervalo.

Para caso das variáveis aleatórias discretas (X) a função distribuição de probabilidade, f(x), fornece, diretamente no eixo y, as probabilidades de ocorrência dos vários valores discretos da variável aleatória no eixo x (x = 1, 2, 3. . .n),

$y = P(X = x) = f(x)$.

Em outras palavras, conhecida a equação de distribuição, basta substituir o valor da variável aleatória na equação, para se obter a probabilidade da ocorrência dela.

Para caso das variáveis aleatórias contínuas (X) a função distribuição de probabilidade, f(x), tem outro significado. Na verdade, ela é o que chamamos de função densidade de probabilidade. Dessa forma, o que ela fornece, no eixo y, são os valores de densidade de probabilidade para os diversos valores da variável aleatória do eixo x.

Antes de prosseguir no raciocínio de como chegamos a valores de probabilidade, no caso das variáveis aleatórias contínuas, lembremos do significado de densidade. Densidade de alguma coisa é o quanto teremos dessa coisa em um determinado volume, área ou comprimento. Então, para podermos calcular uma determinada quantidade de coisa, dada a densidade dela, será necessário definir valores diferentes de zero, de volume, ou de área, ou de comprimento, onde essa coisa está contida. Ou seja, não faz sentido calcular a quantidade de alguma coisa, em volumes, áreas ou comprimentos iguais a zero.

Consideremos o caso da densidade de alguma coisa que varia ao longo de uma linha (eixo x). Essa densidade pode ser expressa da forma de uma função f(x), e pelo racional acima, só é possível calculamos a quantidade dessa coisa em um trecho de linha, com comprimento maior que zero. Em outras palavras, não há como calcular uma quantidade em um ponto da linha.

Fazendo uma analogia, temos uma densidade de probabilidade (função densidade de probabilidade f(x)) que varia ao longo do eixo x. E dessa forma, só possível calcular as probabilidades existentes em trechos e não em pontos do eixo x.

19.4. Distribuições discretas – Distribuição Geométrica – (COMPLETAR!)

A distribuição geométrica é uma distribuição de probabilidade discreta que modela o número de tentativas independentes necessárias para obter o primeiro sucesso em um processo de Bernoulli. Em outras palavras, ela representa o número de falhas consecutivas antes do primeiro sucesso em uma sequência de ensaios de Bernoulli, onde cada ensaio tem uma probabilidade constante de sucesso, denotada por p. Ela é amplamente aplicada em áreas como teoria das filas, análise de séries temporais, modelagem de processos de contagem e em problemas envolvendo tempo até o primeiro evento.

A função de massa de probabilidade (PMF) da distribuição geométrica é dada por:

$$P(X = k) = (1 - p)^{k-1} \times p$$

Onde:

- X é a variável aleatória que representa o número de tentativas necessárias para obter o primeiro sucesso.
- p é a probabilidade de sucesso em uma única tentativa.
- k é o número de tentativas necessárias (começando em 1 para a primeira tentativa).

Principais características da distribuição geométrica:

- Número de Tentativas: A distribuição geométrica pode assumir valores inteiros positivos começando em 1, pois a primeira tentativa é sempre um sucesso.
- Probabilidade de Sucesso: A probabilidade de sucesso em uma única tentativa, denotada por p, deve ser constante para todos os ensaios.
- Decaimento Exponencial: A probabilidade de obter o primeiro sucesso diminui exponencialmente à medida que o número de tentativas aumenta.
- Sem Memória: Uma característica importante da distribuição geométrica é que ela tem a propriedade de falta de memória. Isso significa que a probabilidade de obter o primeiro sucesso na próxima tentativa não é influenciada pelo número de tentativas anteriores.

20. PROB - Distribuições de Probabilidade Contínuas

1. **Distribuição Normal (Gaussiana):**
   - A distribuição normal é amplamente utilizada e é caracterizada pela sua forma de sino. Sua função de densidade de probabilidade é dada por:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

   Onde:

   μ - média da distribuição,

   σ - desvio padrão.

A distribuição normal, também conhecida como distribuição gaussiana, é uma das distribuições estatísticas mais importantes e é caracterizada pela sua forma de sino. Muitos fenômenos na natureza e em várias áreas da ciência e engenharia seguem a distribuição normal. Aqui estão alguns exemplos:

- Alturas Humanas: A altura das pessoas em uma população geralmente segue uma distribuição normal.
- Pontuações em Testes Padrão: As pontuações em testes padronizados, como o Quociente de Inteligência (QI), muitas vezes exibem uma distribuição normal.
- Erro de Medição: Os erros de medição em experimentos científicos, quando somados, tendem a seguir uma distribuição normal, de acordo com o Teorema Central do Limite.
- Peso ao Nascer: O peso dos recém-nascidos em uma população geralmente segue uma distribuição normal.
- Tempo de Resposta em Sistemas Computacionais: O tempo que um sistema de computador leva para responder a uma solicitação pode seguir uma distribuição normal.
- Variação Genética: Muitas características genéticas em uma população, como a cor dos olhos, podem ser modeladas por uma distribuição normal.
- Medições de Pressão Sanguínea: As medições de pressão sanguínea em uma população podem ser aproximadamente modeladas por uma distribuição normal.

- Ruído em Sistemas de Comunicação: O ruído em sistemas de comunicação muitas vezes segue uma distribuição normal.
- Desvios Padrão em Finanças: Os retornos de investimentos financeiros podem ser modelados usando uma distribuição normal, especialmente no contexto do modelo de precificação de ativos financeiros.
- Distribuição de Notas em Exames Padrão: As notas em exames padronizados, quando a população é grande o suficiente, podem se aproximar de uma distribuição normal.

A distribuição normal está relacionada a variáveis aleatórias (v.a.) contínuas. Por conta disso, um dos gráficos que existem para ela é o da função de distribuição de probabilidade ou, função densidade de probabilidade (fdp). Neste gráfico, temos no eixo x os vários valores da v.a (eventos). Mas no eixo y, o que temos são os valores das densidades de probabilidade para os vários pontos x, e não os valores da probabilidade nos pontos.

Com foi visto, pelo fato da distribuição ser normal, relacionada a variáveis aleatórias (v.a.) continuas, não existem, no eixo y, valores de probabilidade associados a valores específicos da v.a., no eixo x. O que existe no eixo y, são valores da "densidade de probabilidade" associadas aos pontos do eixo x (valores da v.a.). Neste caso, são colocados os intervalos cujos extremidades estão os valores da variável aleatória (v.a.) de interesse, que pode ser uma das várias mencionadas no exemplo acima. A probabilidade de encontrarmos valores da v.a. (com uma média **m** e desvio padrão **s**, conhecidos), dentro do intervalo, é dada pela área sob a curva da equação da gaussiana, ou função de densidade de probabilidade, delimitada pelo intervalo.

É importante observar que, embora muitos fenômenos possam ser aproximadamente modelados por uma distribuição normal, nem todos seguem rigorosamente essa distribuição. O uso da distribuição normal é frequentemente justificado pela aplicação do Teorema Central do Limite em situações onde há uma grande quantidade de observações independentes.

20.1. Distribuições Contínuas – Distribuição Exponencial

2. **Distribuição Exponencial:**
   - A distribuição exponencial é comumente usada para modelar o tempo entre eventos em um processo de Poisson. Sua função de densidade de probabilidade é:

     $f(x;\lambda) = \lambda e^{-\lambda x}$

     Onde:

     λ - taxa de ocorrência dos eventos.

Exemplos de fenômenos que seguem a distribuição Exponencial:

A distribuição exponencial é comumente encontrada em muitos fenômenos naturais e processos estocásticos. Aqui estão alguns exemplos de fenômenos que seguem a distribuição exponencial:

- Tempo entre eventos sucessivos: Por exemplo, o tempo entre chegadas de clientes a um caixa em um supermercado, o tempo entre falhas de um sistema mecânico, o tempo entre chamadas telefônicas recebidas em uma central de atendimento.
- Tempo de vida de objetos ou sistemas: O tempo até a falha de um componente eletrônico, o tempo de vida de uma lâmpada, o tempo até a degradação de um material.
- Decaimento radioativo: O tempo que leva para que metade dos átomos de uma substância radioativa se desintegrem segue uma distribuição exponencial.
- Tempo de espera em filas: O tempo que um indivíduo espera em uma fila de supermercado, em um guichê de atendimento ao cliente ou em um semáforo.
- Tempo de resposta em sistemas de computação: O tempo que um processo leva para ser concluído, o tempo entre solicitações de acesso a um servidor, o tempo entre chegadas de pacotes em uma rede.

Esses são apenas alguns exemplos comuns, mas a distribuição exponencial pode ser aplicada em uma variedade de contextos onde o tempo é uma variável importante e os eventos ocorrem de forma independente e aleatória ao longo do tempo.

Então, no gráfico da função de distribuição de probabilidade, no eixo x são colocados os intervalos cujos extremidades estão os valores da variável aleatória de interesse, que pode ser uma das várias mencionadas no exemplo acima, de tempo entre eventos, com uma taxa média **l** de ocorrências, conhecida. A probabilidade da ocorrência de tempos entre

eventos estar em um dado intervalo, é dada pela área sob a curva decrescente da equação de distribuição exponencial, delimitada pelo intervalo.

20.2. Distribuições Contínuas – Distribuição Uniforme

5. **Distribuição Uniforme:**
   - A distribuição uniforme atribui probabilidades iguais a todos os valores em um intervalo específico. Sua função de densidade de probabilidade é constante dentro do intervalo e zero fora dele:

     $f(x; a, b) = \frac{1}{b-a}$

     Onde:

     a - limite inferior do intervalo,

     b - limite superior do intervalo.

A distribuição uniforme é caracterizada pela igual probabilidade de ocorrência para todos os valores dentro de um intervalo específico. Aqui estão alguns exemplos de fenômenos que seguem a distribuição uniforme:

- Lançamento de dados: Quando você lança um dado honesto de seis faces, cada face tem a mesma probabilidade de aparecer, seguindo uma distribuição uniforme.
- Sorteios aleatórios: Se um sorteio é feito de um grupo de números inteiros, onde cada número tem a mesma chance de ser escolhido, então a distribuição dos números sorteados é uniforme.
- Tempo de espera aleatório: Suponha que você esteja aguardando um ônibus em uma parada onde os ônibus chegam a cada 15 minutos. Se você chegar em um momento aleatório durante esse intervalo de 15 minutos, o tempo que você terá que esperar até o próximo ônibus segue uma distribuição uniforme dentro desse intervalo.
- Posição de partículas em um recipiente: Se as partículas estão distribuídas aleatoriamente em um recipiente, sem nenhuma preferência por uma posição específica, então a distribuição das posições das partículas dentro do recipiente pode ser modelada como uma distribuição uniforme.
- Horário de chegada dos clientes em um restaurante: Se o horário de chegada dos clientes em um restaurante é aleatório e não influenciado por fatores externos, então a distribuição dos horários

de chegada pode ser aproximada por uma distribuição uniforme dentro do horário de funcionamento do restaurante.

- Esses exemplos ilustram situações em que a distribuição uniforme é aplicável, onde cada valor dentro de um intervalo tem a mesma probabilidade de ocorrência.

21. PROB - Funções de Densidade de Probabilidade e Funções de Massa de Probabilidade.

Dada uma variável aleatória **X**, **função densidade de probabilidade f(x)** é uma que satisfaz as seguintes condições:

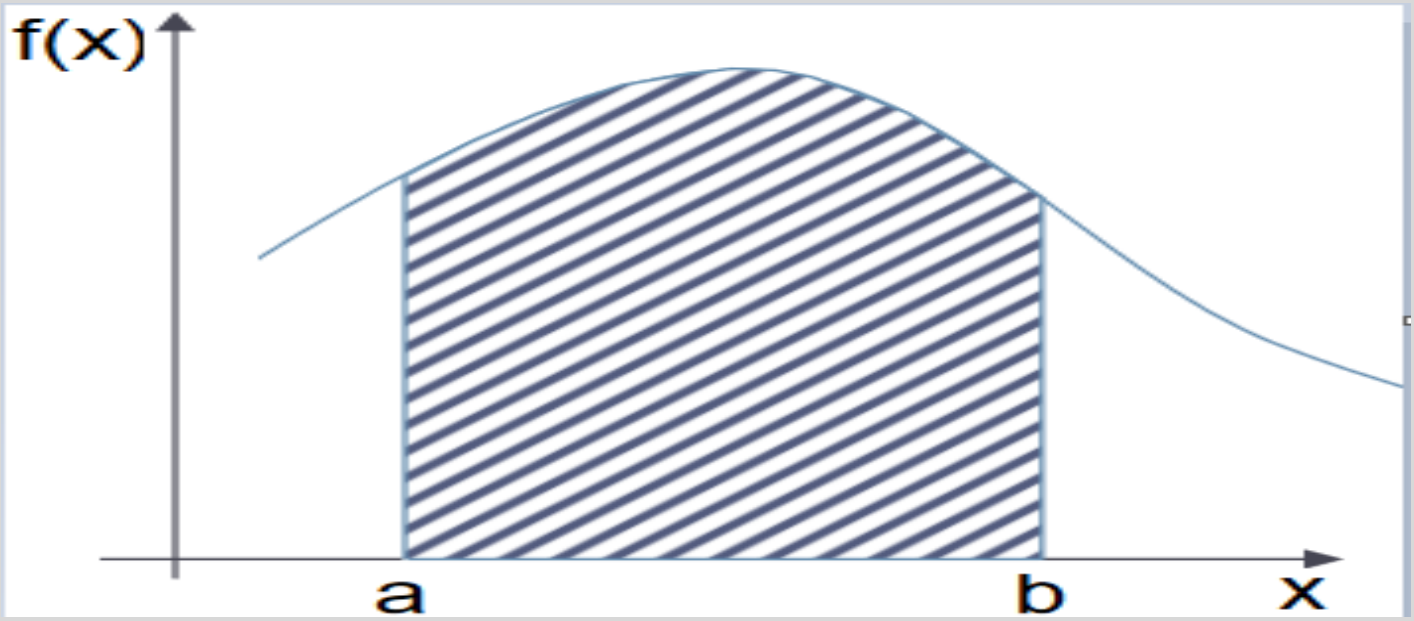
I. $f(x) > 0$ para todo $x \in R_x$

$$\int_{R_x} f(x) = 1$$

Uma **função densidade de probabilidade** é uma função $f(x)$ que satisfaz as seguintes propriedades:

- $f(x) \geq 0$
- $\int f(x)dx = 1$
- Dada uma função $f(x)$ satisfazendo as propriedades acima, então $f(x)$ representa alguma variável aleatória contínua X , de modo que

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$



No caso das variáveis aleatórias contínuas, não há como trabalhar com probabilidades da ocorrência de valores específicos. O que é possível calcular é probabilidade da variável aleatória estar compreendida dentro de um intervalo de valores. Essa probabilidade é dada pelo integral da função f(x) calculada entre os extremos de um intervalo, ou seja, é a área da curva entre os pontos que definem o intervalo.

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

$$P(X = a) = 0$$
$$P(X = b) = 0$$

A probabilidade de ocorrência de um valor específico, em um experimento, vale, matematicamente, sempre 0. No entanto isso não significa que o evento não possa ocorrer. Então, pode-se afirmar o seguinte:

- A probabilidade da ocorrência de um evento impossível é 0;
- Mas se a probabilidade matemática calculada, da ocorrência de um evento, for igual a 0, isso não significa que é impossível que o evento ocorra
- **Função Distribuição de Probabilidade ou de Distribuição Acumulada de Probabilidade**

  A função de distribuição ou acumulada de probabilidade é definida a partir da função de densidade de probabilidade.

  Dada uma variável aleatória contínua X com uma função densidade de probabilidade f(x), a função de distribuição de probabilidade é dada por

  F(x) = P(X < x)

Função de distribuição de probabilidade

Dada uma variável aleatória X, a **função de distribuição** de X é definida por

$$F_X(x) = \mathrm{P}(X \leq x) \qquad \forall x \in \mathbb{R}$$

Def: A **função de distribuição acumulada** $F(x)$ de uma v.a. contínua X é definida para todo x real:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(v)\,dv$$

Onde f(v) é a função densidade de probabilidade, na variável v.

21.1. Propriedades das distribuições (média, variância, momentos).

x

22. PROB - Estatísticas descritivas

22.1. Medidas de tendência central (média, mediana, moda).

x

22.2. Medidas de dispersão (variância, desvio padrão, intervalo).

x

22.3. Medidas de forma (assimetria, curtose).

x

23. PROB - Teoria da amostragem:

23.1. Populações e amostras.

x

23.2. Distribuição de uma estatística amostral.

x

23.3. Amostras e Distribuições Amostrais

x

23.4. Aplicação à Teoria da Confiabilidade

x

23.5. Teorema Central do Limite

TEOREMA CENTRAL DO LIMITE

Se uma variável aleatória X puder ser representada pela soma de quaisquer n variáveis aleatórias independentes, que satisfaçam certas condições gerais, então esta soma, para n suficientemente grande, terá distribuição aproximadamente normal.

Teorema: (Teorema Central do Limite - variáveis aleatórias i.i.d.) Seja $X_1, X_2, \ldots, X_n$ uma sequência de n variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) com $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$. Então, para $S_n = \sum_{i=1}^{n} X_i$, tem-se

$$Z_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

tem uma distribuição aproximada **N(0,1)** na medida em que n se aproxima do infinito. Se F_n é a função de distribuição de Z_n, então

$$\lim_{n\to\infty} \frac{F_n(z)}{\Phi(z)} = 1, \qquad \text{para todo z.}$$

O fato de S_n ser aproximadamente normalmente distribuída quando os termos X_i podem ter qualquer distribuição é a razão básica para a importância da distribuição normal.

Em numerosas aplicações, a variável aleatória considerada pod ser representada como a soma de *n* variáveis aleatórias independentes, algumas das quais podem se dever a erros de medidas, algumas se devem a considerações físicas, entre outras, de modo que a distribuição normal fornece uma boa aproximação.

Em termos mais simples, o TCL nos diz que, mesmo que as variáveis individuais não sejam normalmente distribuídas, a média das amostras dessas variáveis se aproxima de uma distribuição normal conforme o tamanho da amostra aumenta. Isso é crucial em estatística, pois permite inferências sobre a população com base em amostras mesmo quando a distribuição da população não é conhecida.

24. PROB - Estimação pontual e intervalar:

24.1. Estimadores pontuais (método dos momentos, método da máxima verossimilhança).

x

24.2. Intervalos de confiança.

x

24.3. Intervalos de confiança para médias, proporções e variâncias.

x

25. PROB - Testes de hipóteses:

25.1. Formulação de hipóteses nula e alternativa.

x

25.2. Testes de significância.

x

25.3. Erros do Tipo I e do Tipo II.

x

25.4. Testes paramétricos e não paramétricos.

x

26. PROB - Análise de regressão:

26.1. Regressão linear simples e múltipla.

x

26.2. Modelagem de relação entre variáveis.

x

26.3. Assunções e diagnósticos de regressão.

x

27. PROB - Métodos computacionais em estatística:

27.1. Simulação de Monte Carlo.

x

27.2. Métodos de Bootstrap.

x

27.3. Uso de software estatístico (por exemplo, R, Python).

x

28. IND - Introdução à Estatística Indutiva:

A estatística indutiva envolve a aplicação de métodos estatísticos para fazer inferências, generalizações ou previsões sobre uma população, com base em uma amostra dos dados.

As etapas principais compreendidas pela estatística inferencial incluem:

- Formulação da Hipótese:

  Antes de coletar dados, é necessário formular hipóteses sobre a população em estudo. Isso inclui uma hipótese nula (H0) e uma hipótese alternativa (H1), que são testadas estatisticamente para fazer inferências.

- Coleta da Amostra:

  Uma amostra representativa da população é coletada. A qualidade e representatividade da amostra são fundamentais para a validade das inferências.

- Escolha do Teste Estatístico:

  Com base na natureza dos dados e nas hipóteses formuladas, é escolhido um teste estatístico apropriado. Isso pode incluir testes t como o teste t de Student, testes de chi-quadrado, testes F, entre outros.

- Estabelecimento do Nível de Significância:

  Define-se o nível de significância (geralmente representado por α), que indica a probabilidade de rejeitar a hipótese nula quando ela é verdadeira. Um valor comum é 0,05.

- Análise dos Dados e Cálculo do Estatístico de Teste:

  Realiza-se a análise estatística dos dados da amostra, calculando o estatístico de teste apropriado para testar as hipóteses. O resultado é comparado ao valor crítico ou ao p-valor.

- Tomada de Decisão:

  Com base nos resultados do teste estatístico, decide-se se a hipótese nula é rejeitada ou não. Isso implica aceitar a hipótese alternativa ou não ter evidências suficientes para rejeitar a hipótese nula.

- Inferências e Generalizações:

  Se a hipótese nula for rejeitada, fazem-se inferências sobre a população maior com base na amostra. Isso pode incluir estimativas de parâmetros, construção de intervalos de confiança ou previsões.

- Conclusões e Relatório:

  As conclusões derivadas do teste estatístico são resumidas e apresentadas. Isso inclui informações sobre as inferências feitas, limitações do estudo e implicações práticas.

  A estatística inferencial é amplamente utilizada em pesquisas científicas, estudos de mercado, análise de dados empresariais, entre outras áreas, para fazer afirmações sobre populações com base em informações extraídas de amostras.

28.1. Definição de estatística indutiva.

x

28.2. Importância da inferência estatística na tomada de decisões.

x

28.3. Diferença entre população e amostra.

x

29. IND - Distribuições de Probabilidade:

29.1. Distribuições de probabilidade discretas e contínuas (Binomial, Normal, Poisson, etc.).

x

29.2. Propriedades das distribuições de probabilidade.

x

29.3. Teorema do Limite Central e sua importância na inferência.

x

30. IND - Estimação de Parâmetros:

30.1. Estimadores pontuais e intervalares.

x

30.2. Intervalos de confiança.

x

30.3. Métodos de estimação (máxima verossimilhança, método dos momentos, etc.).

x

31. IND - Testes de Hipóteses:

31.1. Formulação de hipóteses nula e alternativa.

x

31.2. Erros do Tipo I e do Tipo II.

x

31.3. Testes paramétricos e não paramétricos.

x

31.4. Testes de comparação de médias, proporções, variâncias, entre outros.

X

32. IND - Testes da Estatística Inferencial

A estatística inferencial envolve a aplicação de métodos estatísticos para fazer inferências sobre uma população com base em uma amostra dos dados. Existem diversos testes estatísticos, cada um projetado para responder a diferentes perguntas ou hipóteses. Abaixo, descreverei alguns testes comuns e exemplos de situações em que são aplicados:

- Teste t de Student:
  - Objetivo: Comparar as médias de duas amostras independentes para determinar se há diferença estatisticamente significativa entre elas.
  - Exemplo: Comparar as médias de notas de dois grupos de estudantes que foram submetidos a diferentes métodos de ensino.

- Teste de ANOVA (Análise de Variância):

- Objetivo: Comparar as médias de três ou mais grupos independentes para verificar se há diferença estatisticamente significativa entre eles.
- Exemplo: Analisar as médias de desempenho de alunos em diferentes escolas para determinar se há diferenças significativas nas abordagens de ensino.

- Teste qui-quadrado

  - Objetivo: Avaliar a independência entre variáveis categóricas em uma tabela de contingência.
  - Exemplo: Investigar se existe uma associação entre gênero (masculino/feminino) e preferência por diferentes gêneros musicais.

- Teste de correlação de Pearson:

  - Objetivo: Avaliar a força e direção da relação linear entre duas variáveis contínuas.
  - Exemplo: Verificar se há uma correlação entre o tempo dedicado ao estudo e as notas obtidas em um exame.

- Regressão linear:

  - Objetivo: Modelar a relação entre uma variável dependente e uma ou mais variáveis independentes.
  - Exemplo: Prever o rendimento acadêmico de um aluno com base no número de horas dedicadas ao estudo semanalmente e na quantidade de horas de sono.

- Teste de Wilcoxon (ou Mann-Whitney U):

  - Objetivo: Comparar duas amostras independentes quando os dados não atendem aos pressupostos do teste t de Student.

- Exemplo: Comparar a pontuação de desempenho de dois grupos de funcionários em uma empresa que não têm distribuição normal.

- Teste de Fisher:

  - Objetivo: Comparar as variâncias de duas amostras independentes.
  - Exemplo: Determinar se as variâncias de dois métodos de produção são estatisticamente diferentes.

32.1. Teste t de Student

Conceitos básicos de probabilidade.
Distribuições de probabilidade.

32.2. Probabilidade:

Conceitos básicos de probabilidade.
Distribuições de probabilidade.

32.3. Comparação de Grupos:

Teste t para amostras independentes e pareadas.
ANOVA (Análise de Variância).

32.4. Regressão Estatística:

Regressão linear simples e múltipla.
Coeficiente de determinação.

32.5. Análise de Correlação:

Coeficiente de correlação de Pearson e Spearman.

32.6. Desenho Experimental:
Princípios de experimentos.
Fatores e níveis.

32.7. Análise de Variância (ANOVA):
ANOVA de um fator.
ANOVA de dois fatores.

32.8. Estatísticas não Paramétricas:
Testes de Wilcoxon.
Testes de Mann-Whitney.

32.9. Modelos de Regressão Avançados:
Regressão logística.
Regressão polinomial.

32.10. Análise de Séries Temporais:
Conceitos básicos.
Modelos ARIMA.

32.11. Estatística Multivariada:
Análise de componentes principais.
Análise de fatores.

32.12. Aplicações Práticas:
Estudo de casos e projetos práticos.
Aplicações em diferentes áreas, como negócios, ciências sociais e saúde.

33. IND - Comparação de Grupos e Análise de Variância (ANOVA):

33.1. Análise de variância de um fator.

x

33.2. Análise de variância de dois fatores.

x

33.3. Testes de comparações múltiplas.

x

34. IND - Correlação e Regressão:

34.1. Correlação de Pearson e Spearman.

x

34.2. Regressão linear simples e múltipla.

x

34.3. Diagnóstico de regressão.

x

35. IND - Modelagem Estatística:

35.1. Modelos lineares generalizados.

x

35.2. Modelos de regressão logística.

x

35.3. Modelos de regressão não linear.

x

36. IND - Validação de Modelos:

36.1. Métodos de validação cruzada.

x

36.2. Bootstrap.

x

37. IND - Análise de Sobrevivência:

37.1. Funções de sobrevivência.

x

37.2. Modelos de risco proporcional de Cox.

x

38. IND - Aplicações e Estudos de Caso:

38.1. Exemplos práticos de aplicação dos conceitos discutidos.

x

38.2. Estudos de caso com análises estatísticas completas.

x

39. IND - Exemplo de uso da Estatística Inferencial

Vamos considerar um exemplo hipotético que envolve a estatística inferencial. Suponha que uma empresa de produção de peças eletrônicas esteja interessada na eficácia de um novo método de fabricação para reduzir a variabilidade nas dimensões de um componente específico.

Aqui estão as etapas do processo:

- Formulação da Hipótese:
  - Hipótese Nula (H0): O novo método de fabricação não tem efeito significativo na redução da variabilidade.
  - Hipótese Alternativa (H1): O novo método de fabricação tem um efeito significativo na redução da variabilidade.

- Coleta da Amostra:

  Uma amostra de componentes é selecionada aleatoriamente do processo de fabricação que utiliza o novo método.

- Escolha do Teste Estatístico:

  Considerando que estamos interessados na variação das dimensões, poderíamos escolher um teste de variância, como o teste F, para comparar as variâncias entre o método antigo e o novo.

- Estabelecimento do Nível de Significância:

  Define-se o nível de significância, por exemplo, $\alpha = 0,05$.

- Análise dos Dados e Cálculo do Estatístico de Teste:

  As dimensões dos componentes na amostra são medidas e os cálculos são realizados para obter o estatístico de teste F.

- Tomada de Decisão:

  Com base no valor crítico do teste F ou no p-valor, decide-se se a hipótese nula é rejeitada.

- Inferências e Generalizações:

  Se a hipótese nula for rejeitada, a empresa pode inferir que o novo método de fabricação tem um efeito significativo na redução da variabilidade nas dimensões dos componentes.

- Conclusões e Relatório:

  As conclusões são resumidas em um relatório, incluindo detalhes sobre o teste estatístico, os resultados, as implicações práticas e as limitações do estudo.

  Neste exemplo, a estatística inferencial é utilizada para fazer inferências sobre a população de componentes com base na amostra selecionada, proporcionando informações sobre a eficácia do novo método de fabricação em termos de redução da variabilidade nas dimensões.

Estimação pontual e intervalos de confiança.
Testes de hipóteses.
Erros tipo I e tipo II.

40. IND - Considerações Éticas e Limitações

40.1. Ética na análise de dados e interpretação dos resultados.

40.2. Limitações dos métodos estatísticos e possíveis vieses.

41. Função Geratriz de Momentos (completar!)

42. Como Identificar a Distribuição de Probabilidade

A partir de um conjunto de dados extraídos de um experimento estatístico, qual é o método usado para identificar a distribuição de probabilidade associada a ele?

Existem várias maneiras de identificar a distribuição de probabilidade associada a um conjunto de dados provenientes de um experimento estatístico. Aqui estão alguns métodos comuns:

(1) Histograma:

Construir um histograma dos dados pode fornecer uma visualização inicial da forma da distribuição. Isso pode ajudar a identificar se a distribuição é simétrica, assimétrica, unimodal ou multimodal.

(2) Função de Distribuição Empírica (ECDF):

A ECDF é uma função que atribui a cada ponto nos dados a probabilidade de que uma observação seja menor ou igual a esse ponto. Plotar a ECDF pode ajudar a comparar visualmente com funções de distribuição teóricas.

(3) Ajuste de Distribuição Paramétrica:

Utilizar métodos estatísticos para ajustar uma distribuição paramétrica (por exemplo, normal, exponencial, log-normal) aos dados. Isso pode ser feito por meio de métodos como o método dos momentos ou o método da máxima verossimilhança.

(4) Testes de Aderência:

Realizar testes de aderência estatística, como o teste de Kolmogorov-Smirnov ou o teste de Anderson-Darling, para comparar os dados com distribuições teóricas específicas. Esses testes podem indicar o quão bem a distribuição teórica se ajusta aos dados.

(5) Gráficos Quantil-Quantil (Q-Q plot):

Um gráfico Q-Q plot compara os quantis observados dos dados com os quantis teóricos de uma distribuição específica. Se os pontos no gráfico se alinharem aproximadamente com uma linha reta, isso sugere que os dados seguem a distribuição teórica.

(6) Bootstrap:

Usar métodos de reamostragem, como o bootstrap, para estimar a distribuição dos parâmetros da população e derivar intervalos de confiança. Isso pode fornecer insights sobre a variabilidade da estimativa dos parâmetros.

Cada método tem suas próprias vantagens e limitações, e a escolha dependerá do contexto específico do experimento e dos dados disponíveis. Em muitos casos, uma combinação de abordagens é usada para obter uma compreensão mais abrangente da distribuição de probabilidade associada aos dados.

42.1. De que maneira é possível verificar se um conjunto de dados, extraído de um experimento específico, obedece a uma distribuição de densidade de probabilidade conhecida?

Verificar se um conjunto de dados segue uma distribuição de densidade de probabilidade conhecida é uma parte fundamental da análise estatística. Existem várias técnicas e ferramentas disponíveis para realizar essa verificação. Aqui estão algumas abordagens comuns:

Gráficos de Probabilidade ou Quantil-Quantil (Q-Q):

Um gráfico Q-Q compara os quantis teóricos da distribuição conhecida com os quantis observados nos dados. Se os pontos no gráfico se aproximarem de uma linha diagonal, isso sugere que os dados seguem a distribuição proposta. Existem também gráficos P-P (probabilidade-probabilidade) que podem ser utilizados.

Histograma e Curva Teórica:

Construa um histograma dos dados e sobreponha a curva teórica da distribuição de densidade de probabilidade conhecida. Se o histograma e a curva teórica se alinharem razoavelmente bem, isso indica uma boa adequação. Ferramentas visuais, como o teste de Kolmogorov-Smirnov, podem ser usadas para quantificar a discrepância entre as distribuições.

Testes Estatísticos de Ajuste:

Existem vários testes estatísticos que podem ser utilizados para verificar o ajuste dos dados a uma distribuição específica. Alguns exemplos incluem o teste de Kolmogorov-Smirnov, teste de Anderson-Darling, teste de qui-quadrado e o teste de Shapiro-Wilk.

Teste de Normalidade:

Se a distribuição de densidade de probabilidade em questão for a normal, testes específicos de normalidade, como o teste de Shapiro-Wilk ou o teste de Anderson-Darling, podem ser aplicados para verificar a normalidade dos dados.

Gráfico de Resíduos:

Se um modelo específico foi ajustado aos dados, você pode examinar um gráfico de resíduos para verificar se eles se comportam de maneira aleatória, indicando uma boa adequação ao modelo.

Testes Visuais e Estatísticos para Outliers:

Avalie visualmente e estatisticamente a presença de outliers nos dados. Outliers podem distorcer a análise de ajuste.

Bootstrap e Simulações Monte Carlo:

Use métodos de resampling como o bootstrap ou simulações Monte Carlo para gerar distribuições amostrais sob a hipótese nula (dados seguindo a distribuição proposta) e compará-las com os dados observados.

Kurtosis e Assimetria:

Calcule a assimetria e a curtose dos dados e compare com os valores esperados para a distribuição proposta. A assimetria e a curtose podem fornecer pistas sobre a adequação da distribuição.

Ao realizar essas verificações, é importante lembrar que, em muitos casos, os dados podem não se ajustar exatamente a uma distribuição teórica. Algumas discrepâncias podem ser aceitáveis, especialmente em conjuntos de dados pequenos. A interpretação deve ser feita considerando o contexto específico do problema em questão.

42.2. De que maneira é possível verificar se um conjunto de dados, extraído de um experimento específico, obedece a uma distribuição de densidade de probabilidade conhecida?

Quais são os passos pare se chegar a um modelo estatístico teórico, que descreve um fenômeno, a partir de um conjunto de dados obtidos?

Desenvolver um modelo estatístico teórico a partir de um conjunto de dados envolve vários passos. Aqui estão os passos comuns que você pode seguir:

(1) Definir o Problema e Objetivos:

Identifique claramente o fenômeno que está sendo estudado e os objetivos da modelagem estatística. Compreender o contexto é fundamental para a escolha adequada do modelo.

(2) Coletar Dados:

Obtenha um conjunto de dados representativo para o fenômeno em estudo. Certifique-se de que os dados sejam relevantes, precisos e suficientes para a análise.

(3) Explorar os Dados:

Realize uma análise exploratória dos dados para compreender a distribuição, identificar padrões, verificar a presença de outliers e entender a relação entre as variáveis. Gráficos e estatísticas descritivas podem ser úteis nesta etapa.

(4) Selecionar o Tipo de Modelo:

Com base na natureza dos dados e nos objetivos da modelagem, escolha o tipo de modelo estatístico apropriado. Isso pode incluir modelos lineares, modelos de regressão, modelos de séries temporais, modelos de machine learning, entre outros.

(5) Dividir os Dados:

Se o conjunto de dados for grande o suficiente, divida-o em conjuntos de treinamento e teste. O conjunto de treinamento será usado para ajustar o modelo, enquanto o conjunto de teste será usado para avaliar o desempenho do modelo.

(6) Ajustar o Modelo:

Use técnicas estatísticas ou algoritmos de aprendizado de máquina para ajustar o modelo aos dados de treinamento. Isso envolve encontrar os parâmetros do modelo que melhor se ajustam aos dados observados.

(7) Validar o Modelo:

Avalie a qualidade do modelo usando dados de teste ou validação. Isso ajuda a verificar se o modelo é capaz de generalizar para dados não vistos.

(8) Interpretar Resultados:

Analise os resultados do modelo e interprete os parâmetros. Considere a significância estatística, a interpretabilidade e a relevância prática.

(9) Ajustar o Modelo (se necessário):

Com base nos resultados da validação, faça ajustes no modelo, se necessário. Isso pode envolver a inclusão/exclusão de variáveis, modificação de parâmetros ou escolha de um modelo diferente.

(10) Documentar e Comunicar:

Documente o modelo, seus resultados e interpretações. Comunique os resultados de forma clara e compreensível para os interessados, usando gráficos, tabelas e explicações.

(11) Implementar o Modelo (se aplicável):

Se o modelo for destinado à aplicação prática, implemente-o no ambiente desejado, garantindo que seja utilizado corretamente.

Lembre-se de que esse processo é iterativo, e é comum revisar e ajustar o modelo conforme necessário ao longo do tempo. A escolha de técnicas específicas dependerá da natureza dos dados e dos objetivos da análise.

43. Como Gauss Desenvolveu a Curva Gaussiana?

A curva gaussiana, ou distribuição normal, não foi "desenvolvida" por Gauss de uma só vez, mas seu trabalho na teoria dos erros e em estatísticas contribuiu significativamente para a compreensão e aplicação da distribuição normal. Gauss desenvolveu a distribuição normal como parte de seu trabalho sobre o método dos mínimos quadrados, que é uma técnica estatística para ajustar uma curva aos dados experimentais.

O método dos mínimos quadrados visa encontrar a melhor curva (por exemplo, uma linha ou uma função) que minimize a soma dos quadrados dos desvios entre os pontos de dados reais e os valores previstos pela curva. Ao aplicar esse método, Gauss percebeu que as distribuições dos erros se ajustavam a uma curva específica, que agora é conhecida como a distribuição normal.

A fórmula da distribuição normal foi desenvolvida de forma mais geral por outros matemáticos ao longo do tempo, mas a contribuição de Gauss para a teoria dos erros e sua compreensão da distribuição normal foram fundamentais para a consolidação desse conceito na estatística e na matemática aplicada.

ESTATÍSTICA DESCRITIVA

(1)DESC-Introdução à Estatística Descritiva:
Definição de estatística descritiva – OK!
Importância da análise descritiva na compreensão dos dados. – OK!

(2)DESC-Medidas de Tendência Central:
Média aritmética. OK!
Mediana. OK!
Moda. OK!
Médias ponderadas. OK!

(3)DESC-Medidas de Dispersão:
Variância. OK!
Desvio padrão. OK!
Amplitude. OK!
Quartis e intervalo interquartil. OK!

(4)DESC-Distribuição de Frequência e Tabelas de Contingência:
Construção de tabelas de frequência. OK!
Histogramas. OK!
Gráficos de barras. OK!
Gráficos de setores. OK!
Tabelas de contingência e frequência cruzada. OK!

(5)DESC-Representações Gráficas:
Box plot. OK!
Gráfico de dispersão. OK!
Gráfico de linha. OK!
Gráfico de densidade. OK!

(6)DESC-Medidas de Assimetria e Curtose:
Coeficiente de assimetria. OK!
Coeficiente de curtose. OK!
Interpretação dos resultados. OK!

(7)DESC-Análise Exploratória de Dados:
Identificação de outliers. OK!
Análise de padrões e tendências. OK!
Detecção de relações entre variáveis. OK!

(8)DESC-Transformação de Dados:
Normalização. OK!
Padronização. OK!
Log-transformação. OK!

(9)DESC-Resumo Numérico e Gráfico para Diferentes Tipos de Variáveis:
Estatísticas descritivas para variáveis categóricas. OK!
Estatísticas descritivas para variáveis contínuas. OK!
Considerações específicas para variáveis qualitativas e quantitativas. OK!

(10)DESC-Aplicações e Interpretação: bbbb
Exemplos práticos de aplicação dos conceitos discutidos.
Interpretação dos resultados da análise descritiva.

(11)DESC - Considerações Éticas e Limitações:
Ética na análise de dados descritivos.
Limitações da análise descritiva e possíveis vieses.

ESTATÍSTICA INDUTIVA

(1)IND-Introdução à Estatística Indutiva:
Definição de estatística indutiva.
Importância da inferência estatística na tomada de decisões.
Diferença entre população e amostra.

(2)IND-Distribuições de Probabilidade:
Distribuições de probabilidade discretas e contínuas (Binomial, Normal, Poisson, etc.).
Propriedades das distribuições de probabilidade.
Teorema do Limite Central e sua importância na inferência.

(3)IND-Estimação de Parâmetros:
Estimadores pontuais e intervalares.
Intervalos de confiança.
Métodos de estimação (máxima verossimilhança, método dos momentos, etc.).

(4)IND-Testes de Hipóteses:
Formulação de hipóteses nula e alternativa.
Erros do Tipo I e do Tipo II.
Testes paramétricos e não paramétricos.
Testes de comparação de médias, proporções, variâncias, entre outros.

(5)IND-Comparação de Grupos e Análise de Variância (ANOVA):
Análise de variância de um fator.
Análise de variância de dois fatores.
Testes de comparações múltiplas.

(6)IND-Correlação e Regressão:
Correlação de Pearson e Spearman.
Regressão linear simples e múltipla.
Diagnóstico de regressão.

(7)IND-Modelagem Estatística:
Modelos lineares generalizados.
Modelos de regressão logística.
Modelos de regressão não linear.

(8)IND-Validação de Modelos:
Métodos de validação cruzada.
Bootstrap.

(9)IND-Análise de Sobrevivência:
Funções de sobrevivência.
Modelos de risco proporcional de Cox.

(10)IND-Aplicações e Estudos de Caso:
Exemplos práticos de aplicação dos conceitos discutidos.
Estudos de caso com análises estatísticas completas.

(11)IND-Considerações Éticas e Limitações:
Ética na análise de dados e interpretação dos resultados.
Limitações dos métodos estatísticos e possíveis vieses.