Guia de Estudos para a prova bimestral de Estatística

Aula 01: Introdução à Estatística

- **Estatística** é aprender sobre um grande grupo pelo exame de dados de alguns de membros. Estatística é a ciência que trata da coleta, organização, análise e interpretação dos dados para a tomada de decisões.
- **Dados**: informações provenientes de observações, contagens, medições ou respostas.
- **População**: coleção completa de todas as medidas (ou dados) a serem consideradas.
- **Censo**: coleção de dados obtidos por toda a população.
- **Amostra**: subcoleção de membros selecionados de uma população. Uma amostra deve ser representativa de uma população de modo que seus dados possam ser usados para tirar conclusões sobre aquela população.

    Técnicas de amostragem:
    - Amostragem aleatória: todos os elementos de uma população têm chances iguais de serem selecionados.
    - Amostragem aleatória simples: cada amostra possível de mesmo tamanho tem a mesma chance de ser selecionada. (por exemplo, cinco pessoas de cada sala).
    - Amostragem estratificada: elementos de uma população são divididos em dois ou mais subconjuntos, chamados de estratos, que compartilham uma característica similar como idade, sexo, etc. Uma amostra é selecionada aleatoriamente de cada um dos estratos. O uso de uma amostra estratificada assegura que cada segmento da população está representado.
    - Amostragem por conglomerado: utilizada quando a população recai em subgrupos que ocorrem naturalmente, cada um tendo características similares. Dividimos a população em grupos (conglomerados) e selecionamos todos os elementos em um ou mais (mas não todos) conglomerados selecionados (porque se não, não é amostragem - é a população). Ao usar esta forma de amostragem, é preciso assegurar que todos os conglomerados tenham características similares.
    - Amostragem de conveniência: consiste somente em membros da população que são fáceis de contatar (não é recomendo pois não garante a imparcialidade e a veracidade dos dados).
- **Parâmetro**: descrição numérica de uma característica populacional. Ou seja, é uma informação numérica real, não algo calculado a partir de uma amostra. É a informação de toda a população.
- **Estatística**: uma estatística é uma descrição numérica de uma característica amostral.
    - Uma estatística amostral pode diferir de uma amostra para outra, enquanto um parâmetro populacional é constante para a população.

- **Estatística descritiva e Estatística inferencial:**

    - A descritiva é o ramo da estatística que envolve a organização, o resumo e a representação de dados. (pegamos os dados que já existem e organizamos eles).
    - A inferencial é o ramo da estatística que envolve o uso de uma amostra para chegar a conclusões sobre uma população. Uma ferramenta básica no estudo da estatística inferencial é a probabilidade. (Pegamos uma parte dos dados e fazemos conclusões sobre o todo a partir de cálculos).

- **Classificação dos dados:**
    - Dados qualitativos consistem em atributos, rótulos ou entradas não numéricas.
    - Dados quantitativos consistem em medidas numéricas ou contagens.

- **Níveis de mensuração:**
    - Dados no nível **nominal** de mensuração são apenas qualitativos. Não é possível fazer cálculos matemáticos com esse nível.
    - Dados no nível **ordinal** de mensuração são qualitativos ou quantitativos. Dados nesse nível podem ser postos em ordem ou classificados, mas as diferenças entre as entradas de dados não têm sentido matemático. (como, por exemplo, um ranking. Por mais que o primeiro do ranking seja classificado como número 01, não é possível fazer cálculos matemáticos com esse 1).
    - Dados no nível **intervalar** de mensuração podem ser ordenados e é possível calcular diferenças que tenham sentido matemático entre as entradas de dados. No nível intervalar, um registro zero é simplesmente uma posição em uma escala, a entrada não é um zero natural. (por exemplo, temperatura ou anos - até podemos fazer subtrações com eles, mas não podemos dividi-los e chegar a resultados plausíveis. Além disso, o ano 0 ou a temperatura 0° não quer dizer que não há ano ou não há temperatura).
    - Dados no nível **racional** de mensuração (mensuração de razão): como os dados no nível intervalar de razão, só que com a propriedade de que um registro de zero é um zero natural. (já aqui, podemos dizer como exemplo a quantidade de gols que um time fez no jogo de futebol. 0 gols significa nenhum gol). Em dados no nível de mensuração de razão, podemos determinar se um dado é múltiplo do outro.

- **Diferença entre um estudo observacional e um experimento:**
    - Um estudo observacional se caracteriza pela **permanência do ambiente exatamente como ele está**. O pesquisador observa e mede as características de interesse de parte da população, mas não muda as condições existentes.
    - Um experimento se caracteriza pela **interferência no ambiente no qual o estudo está sendo feito.** Um tratamento é aplicado em uma parte da população, chamada de grupo de tratamento. Outro grupo, o grupo de controle, não recebe tratamento. As respostas são comparadas e estudadas.

- **Coleta de dados:**

    - Simulação: uso de um modelo matemático ou físico para reproduzir as condições de uma situação ou processo.
    - Pesquisa: investigação de uma ou mais características de uma população - mais frequentemente, as pesquisas são conduzidas com pessoas, por meio de entrevistas.
- **Planejamento experimental:**
    - **Elementos-chave: controle, aleatorização e reaplicação.**
    - Variável de confusão: ocorre quando um pesquisador não pode distinguir um ou mais fatores que causaram os efeitos provocados sobre a variável em estudo, gerando confusão. (por exemplo, o fator que causou a gripe na população foi realmente o vírus ou só uma frente beeem fria?)
    - Efeito placebo: ocorre quando um indivíduo reage favoravelmente a um tratamento, quando na verdade, recebeu um placebo (a reação é psicológica). Para evitar o efeito placebo, utilizamos os experimentos cegos ou duplo cegos:
        - Experimento cego (ou cegamento): técnica no qual o indivíduo não sabe se está recebendo um tratamento ou um placebo. (para não ter como o paciente reagir psicologicamente).
        - Experimento duplo-cego: nem o pesquisador nem os pacientes sabem quem está recebendo placebos e quem está realmente recebendo o tratamento
    - Aleatorização: processo de se designar indivíduos aleatoriamente para diferentes grupos de tratamento.
        - Planejamento completamente aleatorizado: os indivíduos são designados para diferentes grupos de tratamento por meio de seleção aleatória.
        - Planejamento em blocos aleatorizados: o pesquisador separa os indivíduos com características similares em blocos e, então, dentro de cada bloco, designa-os aleatoriamente para os grupos. (exemplo, dividimos em mulheres e homens e, depois, os grupos são aleatórios - para conseguirmos fazer uma análise melhor).
    - Replicação: É a repetição de um experimento sob condições iguais ou semelhantes.

Aula 02: Estatística Descritiva 01

- **Distribuição de frequência:** agrupamento de informações extensas em classes (intervalos) para melhor análise dos dados. Uma distribuição de frequência é uma tabela que mostra classes ou intervalos dos valores com a contagem do número de ocorrências em cada classe ou intervalo. A frequência de uma classe é o número de ocorrências de dados na classe (lembrando que as classes não se sobrepõem).
- **Amplitude da classe:** distância entre os limites inferiores (ou superiores) de classes consecutivas;
- **Amplitude da distribuição:** diferença entre os valores máximo e mínimo;
- **Ponto médio:** soma dos limites inferior e superior da classe dividida por dois. O ponto médio é, às vezes, chamado de marca da classe ou representante da classe.

- **Frequência relativa:** fração ou proporção de dados que está nessa classe. Para calcular a frequência relativa de uma classe, dividimos a frequência pelo tamanho da amostra (e, se quisermos em porcentagem, multiplicamos por cem).
- **Frequência acumulada**: soma das frequências dessa classe com todas as anteriores. A frequência acumulada da última classe é igual ao tamanho *n* da amostra.

—-------- gráficos —---------
(pegar códigos) - aulas práticas e de gráficos 2 e 3

Aula 04: Estatística Descritiva 03

- **Medida de tendência central**: é um valor que representa uma observação central de um conjunto de dados (ferramentas para análise de dados).
    - **Média**: soma dos valores dos dados dividida pelo número de observações.

      média populacional: $\mu = \Sigma x / N$ (a média populacional "mi" é a soma de todos os dados x dividida pelo número de observações em uma população N)

      média amostral = x linha = $\Sigma x / n$ (a média amostral "x linha" é a soma de todos os dados da amostra x dividida pelo número de observações em uma amostra n)

    - **Média ponderada:** é a média de um conjunto de dados cujos valores têm pesos variados (usado para notas acadêmicas, feedbacks ou questões financeiras, por exemplo).

      média ponderada: x linha = $\Sigma (x * w) / \Sigma w$ (a média ponderada "x linha" é os dados vezes seus respectivos pesos, tudo isso somado e dividido pela soma dos pesos).

    - **Média de uma distribuição de frequência para uma amostra:** a distribuição de frequência não é dividir os dados de um conjunto em classes e colocar a frequência deles? Então, se quiséssemos achar um valor que representasse a grande maioria desses dados, o que faríamos?

      média de uma distribuição de frequência para uma amostra:
      x linha = $\Sigma (x * f) / \Sigma f$ (o "x linha" é igual à multiplicação do ponto médio de uma classe pela frequência dela - fazemos isso com todas as classes e somamos - dividido pelo tamanho da amostra - soma de todas as frequências)

➔ **Outliers:** A nossa média pode ser afetada por números muito discrepantes, pois ela conta como valor e pode aumentar ou diminuir além da conta o "valor representativo". Esses números que são muito diferentes do resto do conjunto são chamados de **outliers**. Alguns outliers podem ocorrer por falhas na coleta de dados.

Logo, as conclusões tomadas com base em um conjunto de dados que contém outliers podem ser falhas.
Uma forma de identificar outliers é com a amplitude interquartil, vista na aula 06.

- **Mediana:** o valor que está no meio de um conjunto de dados quando o conjunto está ordenado. (diferentemente da média, que pega um valor que "representaria" todos os valores, a mediana pega um valor real do conjunto que seria o "elemento do meio").

  Quando o conjunto de dados tem um número ímpar de observações, a mediana é o elemento do meio.
  Quando o conjunto de dados tem um número par de observações, a mediana é a média dos dois elementos que ocupam as posições centrais.

- **Moda:** é o valor do conjunto de dados que ocorre com mais frequência. (então, se houverem cinco observações e três delas forem, sei lá, "sorvete", então "sorvete" é a moda, pois é o valor que mais aparece no conjunto).

  Quando nenhum dado se repete, o conjunto não tem moda. Quando dois valores ocorrem com a mesma frequência, cada um é uma moda e o conjunto é chamado de bimodal.

➔ Regra típica de arredondamento para estatística: empregar sempre uma casa decimal a mais do que o conjunto de dados está empregando (então se o conjunto de dados apresentar números inteiros, empregar uma casa decimal e assim por diante).

- **Formas das distribuições:**
  - **Simétrica:** quando uma linha vertical pode ser desenhada pelo meio do seu gráfico da distribuição e as metades resultantes são imagens espelhadas ou muito parecidas. Quando a distribuição é simétrica e unimodal, **a média, a mediana e a moda são iguais.**
  - **Uniforme (ou retangular):** quando todos os valores ou classes na distribuição têm frequências iguais ou aproximadamente iguais. Uma distribuição uniforme também é simétrica (a média e a mediana são iguais).
  - **Assimétrica:** quando a "cauda" do gráfico se alonga mais em um dos lados. Uma distribuição é assimétrica à esquerda (assimetria negativa - existem menos pessoas acima da média) quando sua cauda se estende para a esquerda **(geralmente média<mediana<moda)**, e assimétrica à direita (assimetria positiva - existem mais pessoas acima da média) quando sua cauda se estende para a direita **(geralmente média>mediana>moda)**.

Aula 05: Estatística Descritiva 04

- **Medidas de variação:** valor que representa a dispersão em um conjunto de dados.
  - **Amplitude:** é a diferença entre o valor máximo e o valor mínimo. Para encontrar a amplitude, os dados precisam ser quantitativos (já que não é possível subtrair dados qualitativos).

- **Desvio de um valor x em uma população:** é a diferença entre o valor x e a média μ do conjunto de dados (o quanto esse valor desvia do padrão).

  desvio de x = x - μ

  É importante lembrar que a soma de todos os desvios para qualquer conjunto de dados é zero (porque, pensa, se você tem um conjunto e faz uma média, praticamente todos os números estarão um pouco acima ou um pouco abaixo dessa média - visto que é um número representativo. Então, se você calcular o quanto cada valor se desvia dessa média e depois somar todos os desvios, essa soma vai resultar que, no final, esse número é representativo desse conjunto). Por isso, se calcula a soma dos quadrados dos desvios, SSx.

- **Variância populacional:** é a média dos quadrados dos desvios (visto que não conseguimos tirar nada da soma dos próprios desvios). Como a unidade de medida é ao quadrado, geralmente utilizamos o desvio padrão populacional.

  variância populacional: $\boldsymbol{\sigma}^2 = \Sigma (x - \mu)^2 / N$ (a variância populacional "sigma minúsculo) é a soma de todos os quadrados dos desvios dividida pelo número de observações de uma população.

- **Desvio padrão populacional:** raiz quadrada da variância populacional para conseguirmos uma unidade de medida manipulável. **Mede a variação dos dados em relação à média** (para conseguirmos uma análise mais completa). É sempre maior ou igual a 0 (visto que o máximo que vai acontecer é todos os dados serem iguais à média - não há desvio negativo). Quanto mais dispersos os valores, maior o desvio padrão.

  desvio padrão populacional: $\boldsymbol{\sigma} = \sqrt{\boldsymbol{\sigma}^2} = \sqrt{\Sigma (x - \mu)^2 / N}$

- **Variância amostral e desvio padrão amostral:** difere um pouco da variância populacional e desvio padrão populacional. Variância amostral é a média dos quadrados dos desvios, só que a média é a média amostral e é dividido pelo número de observações da amostra menos 1 (pelos cálculos de não enviesamento).

  variância amostral: $s^2 = \Sigma (x - \text{x linha})^2 / n - 1$

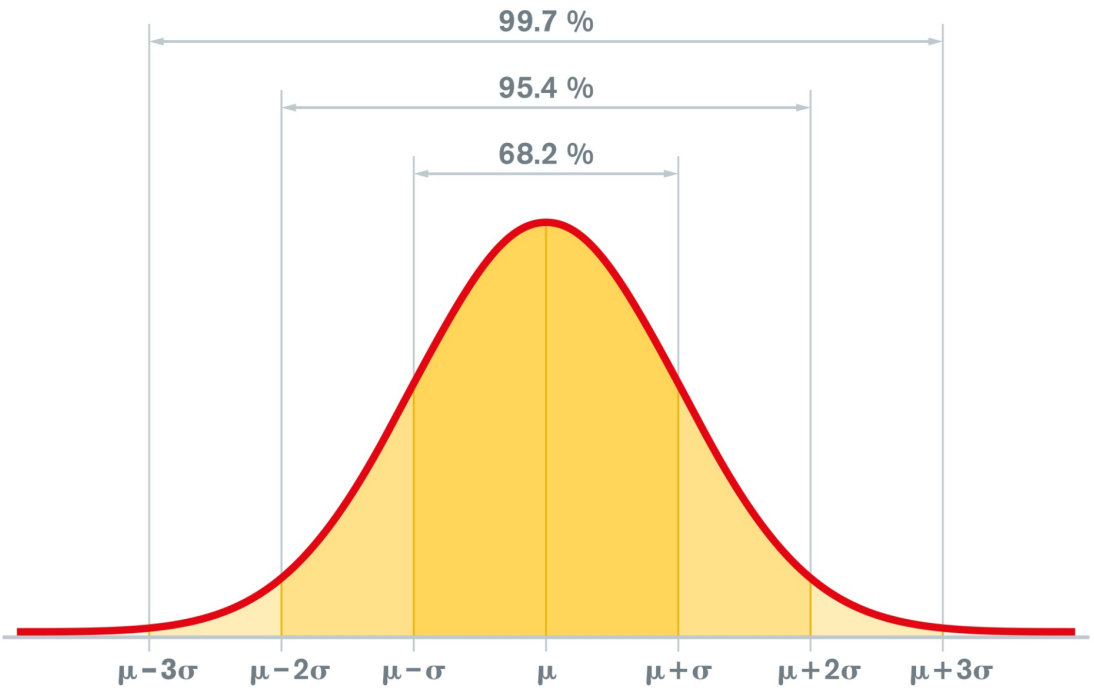
  desvio padrão amostral: $s = \sqrt{s^2} = \sqrt{\Sigma (x - \text{x linha})^2 / n - 1}$

Valores de dados que se encontram a mais ou menos dois desvios padrão da média:	dados incomuns

Valores de dados que se encontram a mais de três desvios padrão da média:	dados muito incomuns

valores incomuns e muito incomuns têm uma influência maior no desvio padrão do que aqueles que estão mais próximos da média (da mesma forma que os outliers influenciam no cálculo da média).

- **Regra Empírica (ou Regra 68-95,99,7):** para conjuntos que são aproximadamente simétricos e possuem formato de sino (ou seja, grande parte dos conjuntos de dados da vida real), o desvio padrão possui essas características:



- **Coeficiente de variação:** utilizamos esse coeficiente que mede a variação de dados com relação à média em porcentagem, para que possamos comparar conjuntos de dados com unidades de medida diferentes ou médias diferentes.

  coeficiente de variação de uma população: CV= σ/μ * 100% (o desvio padrão populacional dividido pela média vezes 100)

  coeficiente de variação de uma amostra: CV= s/ x linha * 100% (o desvio padrão amostral dividido pela média amostral vezes 100)

Aula 06: Estatística Descritiva 05

- **Separatrizes:** números que dividem um conjunto de dados em partes iguais (com o mesmo número de elementos. Precisamos investigar as separatrizes para especificar a posição de um elemento dentro de um conjunto de dados.

  Um exemplo de separatriz é a mediana, pois divide o conjunto em dois subconjuntos com a mesma quantidade de valores.

  - **Quartis:** dividem o conjunto de dados ordenados em quatro partes iguais. O **primeiro quartil** (Q_1) é um valor que marca aproximadamente 25% (¼) dos dados. O **segundo quartil** (Q_2 - que é o mesmo que a mediana) é um valor que marca aproximadamente a metade (50%) dos dados. O **terceiro quartil** (Q_3) é um valor que marca aproximadamente 75% (¾) dos dados. Logo, entre o primeiro e o terceiro quartil, existem aproximadamente metade dos dados (**amplitude interquartil** - amplitude da porção central).

    O primeiro quartil é a mediana entre o menor valor dos dados e a mediana do conjunto.
    O segundo quartil é a mediana do conjunto.
    O terceiro quartil é a mediana entre a mediana do conjunto de dados e o maior valor dos dados.

    Uma forma de identificar outliers é com a amplitude interquartil:
    - multiplicamos a AIQ (amplitude interquartil) por 1,5;
    - fazemos Q_1 - esse valor. qualquer valor abaixo disso é um outlier.
    - fazemos Q_3 + esse valor. qualquer valor acima disso é um outlier.

- **Diagrama de caixa-e-bigode (boxplots) para aplicação dos quartis:** Esse diagrama é de análise exploratória que destaca características importantes de um conjunto de dados.

  - **Decis:** divide o conjunto de dados em 10 partes iguais (D_1 a D_9);
  - **Percentis:** divide o conjunto de dados em 100 partes iguais (P_1 a P_{99}).
    para encontrar o percentil correspondente a um valor específico x, divida a quantidade de elementos menores do que x pelo número total de elementos e multiplique por 100, arredondando o resultado para o valor inteiro mais próximo.

- **Escore padrão:** quantos desvios o valor x se encontra longe da média.

  escore padrão (ou escore-z): $z = x - \mu / \sigma$

  Se z for menor que 0, x é menor que a média;
  Se z for igual a 0, x é igual à média;
  Se z for maior que 0, x é maior que a média.

Aula 07: Probabilidade

- **Experimento probabilístico:** ação ou tentativa sujeita à lei do acaso, pela qual resultados específicos são obtidos.
    - **Resultado:** produto de uma única tentativa em um experimento probabilístico.
    - **Espaço amostral:** conjunto de todos os experimentos possíveis em um experimento probabilístico.
    - **Evento:** subconjunto do espaço amostral (um ou mais resultados).
        - **Evento simples:** evento com condição que leva somente a um resultado possível (ex.: tirar 2 no dado).
        - **Evento não simples:** evento com condição que leva a mais de um resultado possível (ex: tirar um número par no dado).

    - Podemos utilizar o **diagrama de árvore** para determinar a quantidade de resultados possíveis (espaço amostral).
    - **O princípio fundamental da contagem:** para definir o espaço amostral de eventos em sequência. O número de maneiras que dois eventos podem ocorrer em sequência é m*n (sendo *m* a quantidade de maneiras possíveis que o primeiro evento ocorre e *n* a quantidade de maneiras possíveis que o segundo evento ocorre).
    - **Tipos de probabilidade:** P(E) significa “a probabilidade do evento E”.
        - **Probabilidade teórica (ou clássica):** quando cada resultado em um espaço amostral tem a mesma possibilidade de ocorrer, sendo:

          P(E) = número de resultados no evento E / número total de resultados no espaço amostral.

        - **Probabilidade empírica (ou estatística):** baseada em observações obtidas de experimentos probabilísticos.Quando um experimento é repetido muitas vezes, são formados padrões regulares que permitem encontrar a probabilidade empírica. A probabilidade empírica de um evento E se dá por:

          P(E) = frequência total do evento E no experimento probabilístico (f) / frequência total do experimento probabilístico (n)

        - **Lei dos números grandes:** conforme um experimento é repetido muitas muitas vezes, a probabilidade empírica de um evento tende a se aproximar da sua probabilidade teórica (real).
        - **Probabilidade subjetiva:** probabilidade resultante de conjecturas e de estimativas por intuição.
        - **Regra da amplitude das probabilidades:** $0 \leqslant P(E) \leqslant 1$
        - **Complemento do evento E:** conjunto de todos os resultados em um espaço amostral que não estão incluídos no espaço amostral (E’). Quando sabemos a probabilidade de um evento E, podemos calcular a probabilidade do complemento de E.

          P(E) + P(E’) = 1

Aula 08: Probabilidade Condicional

- **Probabilidade condicional:** é a probabilidade de um evento ocorrer, dado que outro evento já tenha ocorrido. Denotação: P(B|A) - “probabilidade de B, dado A”.
- **Eventos independentes:** quando a ocorrência de um evento não afeta a probabilidade de outro.
- **Eventos dependentes:** quando a ocorrência de um evento afeta a probabilidade de outro.
    - Para determinar se A e B são dependentes ou independentes, calculamos a probabilidade do evento B - P(B). Depois calculamos a probabilidade do evento B, dado A - P(B|A). Se os valores forem iguais, os eventos são independentes. Se os valores forem diferentes, os valores são diferentes.
- **A regra da multiplicação para a probabilidade de A e B:** A probabilidade de dois eventos ocorrerem em sequência:
    - Se A e B forem independentes: P(A e B) = P(A) * P(B)
    - Se A e B forem dependentes: P(A) * P(B|A)

conceitos	**para quê usamos na prática**
amplitude	se as médias dos conjuntos forem as mesmas, podemos saber o qual tem dados mais dispersos
variância e desvio padrão	saber o quão dispersos estão os dados
coeficiente de variação	comparar entre dois conjuntos com unidades diferentes qual tem maior dispersão de dados
Escore padrão	Usado para identificar valores incomuns no conjunto que dados que tem formato aproximado de sino.
princípio fundamental da contagem	para definir o espaço amostral de eventos em sequência