O QUADRO GERAL DA ESTATÍSTICA

A estatística é uma ciência multidisciplinar dedicada ao estudo e à aplicação de procedimentos de recolha e técnicas de análise de dados, com o objetivo de responder a questões de investigação. A resposta a essas questões depende do grau em que os dados fornecem evidência empírica suficiente para fundamentar decisões informadas sobre a melhor resposta à questão de investigação, minimizando o risco de erro em comparação com outras decisões alternativas.

De forma simplificada, a estatística pode ser definida como um **método para responder perguntas de investigação** com base no uso de dados a larga escala. Por método se entende uma abordagem prática com suporte teórico, que inclui uma extensa variedade de técnicas. Por larga escala se entende um alto número unidades de observações (*large N*), tais como eleitores de um país, horas passadas frente aos ecrãs, ou variações no preços de bens e serviçoes. baseados em um número reduzido de observações (*small N*) ou em um único caso (*case study*), mais comuns nas abordagens qualitativas s (Gerring XXX)

A **pergunta de investigação** representa o ponto de partida essencial para qualquer estudo acadêmico ou científico, servindo como um guia para a formulação de hipóteses e a escolha de métodos de análise. Ela consiste no enquadramento de um problema — de natureza teórica, empírica ou social — de forma clara, delimitada e operacionalizável, estruturando a relação entre uma ou mais variáveis.[1]

Pergunta de investigação exemplo

- *Qual é o impacto da situação laboral no comportamento político dos jovens portugueses?*

Uma característica método estatístico é que, guiado pela pergunta de investigação, que permite fazer uma cobertura integral do ciclo da análise empírica, seguindo as fases que identificamos a continuação.

1. Delimitação do Fenómeno de Estudo: População, Variáveis

Um primeiro passo na analise estatística é a identificação da população de interesse, isto é, do universo de casos que pretendem ser estudados. Para além de pessoas, este conjunto pode referir a

[1] As perguntas de investigação são clasificadas segundo o número e relação dessas variávies. Por exemplo, para Cícero et al. (20xx), distinguem 4 tipos dependendo de dois fatores: (1) se envolvem ou não relações entre variáveis e (2) se são formuladas a partir de teorias prévias.

unidades de qualquer tipo tais como organizações, comunidades, países, discursos ou protestos sociais. E, mesmo que não se discuta muito de modo explícito, trata-se sempre de uma construção teórica onde os limites podem não ser evidentes. Por exemplo, qual é a idade para ser considerado um "jovem português"? E os jovens portugueses no estrangeiro devem ser incluídos ou excluídos? E os jovens estrangeiros morando em Portugal? As respostas para estas perguntas não devem (não podem) ser verdadeiras ou falsas mas sim explícitas e justificadas. Note-se também que um mesmo indivíduo pode ser simultaneamente membro de diferentes populações. Um jovem também pode ser parte dos grupo "mulheres", "classe media", "criado na periferia", "com educação alta"; por enquanto Portugal é ao mesmo tempo um indivíduo das populações "país europeio", "sistemas semi-presidenciais", ou "culturas lusófonas". A definição da população é um momento analítico chave da análise, com implicâncias cruciais nos processos de análise e fiabilidade dos resultados obtidos.[2]

População da pergunta de investigação exemplo

- *Os jovens portugueses*

A seguir, é preciso identificar as variáveis da população que sejam pertinentes para dar uma resposta à pergunta de investigação. Perguntas bem formuladas costumam conter esta informação de modo explícito. A "condição laboral" e "comportamento político" dos jovens portugueses seriam as duas características chaves para lidar com a pergunta de investigação exemplo.

Variáveis chave da pergunta investigação exemplo

- *Condição laboral dos jovens portugueses*
- *Comportamento político dos jovens portugueses*

Outras variáveis, não presentes pergunta mas sim na literatura relacionada, poderia ser o género, a localidade de residência, ou o nível educativo. Estas terceiras variáveis costumam ser incluídas nos modelos estatísticas como elementos de controlo.

2. Recolha dos dados: Amostra, Indicadores Empíricos e Estatísticas

Um segundo passo do método estatístico é a produção de dados, que inclui dois procedimentos: a seleção de um subconjunto dos indivíduos da população estudada (amostra), e a escolha dos indicadores empíricos para a observação das variáveis de interesse. Embora o objetivo final da análise estatística seja obter conclusões sobre a população, abrangê-la integralmente pode ser difícil ou impraticável, seja por sua extensão, volatilidade ou falta de acessibilidade. Além disso, em muitos casos, não é necessário, pois a informação pode tornar-se redundante. A solução para esse desafio é a construção de uma amostra estatística, que deve seguir critérios específicos para garantir sua representatividade. Basicamente, uma amostra é válida so se pode ser utilizada como uma boa representação da totalidade observada. Esta se diferencia da **evidencia anedótica** resultante da

[2] (qualitativo ao contrario: a partir do caso se pensa o universo. Um mesmo caso poder pertencer a diferentes populações) concept building

observação de casos isolados ("o meu tio fumou toda a vida e chegou ao 98 anos em perfeito estado"). Se diferencia dos **censos**, que observam à totalidade da população. Como se mencionou, isto costuma ser difícil de realizar por diferentes razões mas também pode ser desnecessário: para concluir quanto de sal a sopa tem, basta testar uma colher depois de uma boa mexida. Uma boa amostra estatística tem como principal preocupação fugir dos **enviesamentos** da sua composição interna. Uma amostra de jovens portugueses integrada exclusivamente por estudantes universitários estaria enviesada por estar restrita a um subgrupo específico, ignorando informação para todo o resto. Isto costuma passar com as nomeadas **amostras de conveniência**, construídas com indivíduos de fácil acesso para o investigador. O enviesamento nem sempre é tão claro de identificar nem de gerir. Contactar os jovens por telefone fixo poderia ter uma boa estratégia 20 anos atrás, hoje recolheríamos muitas poucas observações. Mas o número de observações também não é suficiente para garantir qualidade e pertinência da informação obtida: a utilização do whatsapp poderia conseguir aceso a um altíssimo número de indivíduos, mas está igualmente esquecendo sectores importantes da população alvo que não usa o Whatsapp. O enviesamento pode ser difícil e até impossível de eliminar completamente, mas é um aspeto que sempre deve ser considerado. E, para compensar, amostras de conveniência podem ser amostras apropriadas se acompanhadas de justificação correspondente dependendo do tipo de pergunta de investigação colocada. Por exemplo, uma amostra composta só de estudantes universitários tem o potencial para ser uma boa amostra da população jovens portugueses estudantes universitários.

A amostra estatística baseia-se no princípio da aleatoriedade. Um exemplo intuitivo desse processo é o funcionamento de um sorteio: imaginemos uma caixa contendo papéis numerados, cada um representando um indivíduo de uma determinada população. Após misturar bem os papéis, sorteia-se uma porção deles, formando assim a amostra. Esse procedimento ilustra a técnica da **amostragem aleatória simples** (*simple random sampling*), que é adequada para populações de tamanho conhecido e para as quais dispomos de uma lista completa de todos os seus membros. No entanto, quando não temos acesso a essa lista, mas possuímos informações sobre características importantes que diferenciam os indivíduos dentro da população, pode ser mais apropriada a **amostragem estratificada** (*stratified sampling*). Nesse caso, a população é primeiro dividida em estratos homogêneos, conforme um critério relevante (por exemplo, zonas rurais e urbanas). Em seguida, dentro de cada estrato, os indivíduos são selecionados aleatoriamente. Já para populações onde há grande variabilidade entre grupos vizinhos—por exemplo, quarteirões com casas grandes e pequenas—pode ser útil a **amostragem por conglomerados** (*cluster sampling*). Aqui, a população é dividida em grupos heterogêneos sem seguir um critério específico, e apenas alguns desses grupos são sorteados aleatoriamente para compor a amostra.

O método estatístico depende fortemente das propriedades da amostra, para a qual são estabelecidos procedimentos e requisitos específicos de construção. Os mais importantes são o tamanho da amostra (N) e sua composição interna, garantindo a ausência de vieses.

[Figura amostras? ver openIntro]

Após a seleção da amostra, é necessário identificar os **indicadores empíricos** adequados para observar as variáveis de análise. Com base na literatura existente, a 'condição laboral' poderia ser observada através de indicadores como o status de emprego (se o indivíduo está empregado ou não), enquanto o 'comportamento político' poderia ser registrado pela sua participação eleitoral. A escolha desses indicadores faz parte do processo de operacionalização dos conceitos, que envolve

uma redução na escala de abstração. Trata-se, mais uma vez, de uma decisão do investigador, que requer uma justificação teórica explícita.

Indicadores empíricos chave da pergunta investigação exemplo

- *Salários dos jovens portugueses*
- *Partido votado pelos jovens portugueses*

Os indicadores empíricos determinam os modos pelos quais as variáveis registram as diferenças entre os indivíduos de uma população. Por exemplo, o salário pode ser registrado com números (variando entre 0 e vários milhares de euros), enquanto o tipo de partido votado pode ser representado por um número reduzido de categorias (esquerda/centro/direita). O modo de medir essa variação dependerá em grande parte da característica observada, mas também costuma ser uma escolha 'fundamentada' do investigador. Essa variação introduz um dos aspectos mais característicos da análise estatística: os tipos de variáveis, conforme sua escala de medição. As denominações utilizadas para classificar esses tipos podem variar entre disciplinas, áreas de estudo e pacotes estatísticos, mas o seu significado permanece similar.

Uma primeira distinção é entre variáveis numéricas e categóricas. As **variáveis numéricas** representam valores que podem ser quantificados (ou seja, há variação entre os números) e podem ser tratadas com operações aritméticas, como somar, subtrair, multiplicar ou dividir. Exemplos disso incluem a quantidade de cigarros consumidos, o número de votos obtidos, o número de filhos, a inflação mensal ou o montante de euros gastos. Por exemplo, se o indivíduo A gastou 5 euros e o indivíduo B gastou 3 euros, o total de gastos dos dois é 8 euros.

Já as **variáveis categóricas** representam tipos ou classes qualitativamente distintas entre si (ou seja, há variação entre as categorias), mas não podem ser manipuladas por operações aritméticas. Exemplos de variáveis categóricas são distinções como branco/preto, português/espanhol ou maçãs/pêras. Por exemplo, se o indivíduo A comeu 5 maçãs e o indivíduo B comeu 3 pêras, não podemos somar as maçãs e as pêras, pois elas pertencem a categorias diferentes. O total de frutas consumidas seria de 5 maçãs e 3 pêras, mas sem que possamos tratá-las como se fossem a mesma coisa.

Embora essa distinção entre variáveis quantitativas e categóricas seja suficiente para entender e aplicar a maioria dos conceitos estatísticos, como explicado em *OpenIntro:6*, é importante também ter em mente algumas subdivisões mais específicas.

Num segundo nível, as variáveis numéricas podem ser distinguidas em discretas e contínuas. As ***variáveis numéricas discretas*** registam unidades claramente delimitadas, com espaços constantes e conhecidos e representáveis com símbolos numéricos inteiros (e.g., 1-2-3-4...), tais como número de filhos ou de legisladores duma assembleia: um partido político poder ter 20 ou 21 legisladores próprios mas não 20,5. Entretanto, as ***variáveis numéricas contínuas*** têm todas as propriedades das discretas, mas os seus espaços são monotonicamente crescentes, podendo ser representados por qualquer valor num intervalo por meios de símbolos numéricos decimais (1,0-1,1-1,2-...) tais como a idade ou a temperatura. Logo, o valor zero de uma numérica pode referir simplesmente a alguma convenção (0 grado de temperatura) ou pode estar indicando explicitamente a ausência de uma quantidade (0 filhos, 0 legisladores). No segunda caso, tratar-se-ia de uma **variável numérica *de***

razão por que é possível dizer que uma quantidade é maior que outra em X vezes. Por exemplo, 60 legisladores é 2 vezes maior do que 30 legisladores.

Pela parte das variáveis categóricas, dois subtipos distinções adicionais devem ser consideradas. As variáveis categóricas que podem ser ordenadas seguindo um critério de hierarquia (por exemplo: baixo, medio, alto) são denominados **variáveis ordinais**. Em tanto que variáveis de 2 categorias são frequentemente referidas como **variáveis dicotómicas**. Finalmente, tenha-se presente que, para facilitar a codificação e o tratamento dos dados, é frequente utilizar símbolos numéricos para referir as categorias (0: homens, 1=mulheres; 1=pouco; 2: algo; 3: muito). Estes números são simples códigos de referência que não podem ser tratados aritmeticamente.

Os dados empíricos são recolhidos e organizados em matrizes de dados compostas de linhas e colunas. São as **bases de dados** que, para além de serem um componente vital da análise estatística, fornecem uma das ilustrações mais claras dos diferentes elementos de uma análise empírica qualquer. As linhas da matriz representam as **unidades de observação**. O número total delas indica o **tamanho da amostra** (N). As colunas: representam as **variáveis**. Toda base de dados é (ou deveria ser) acompanhada de um livro de códigos, o qual especifica a escala de medição das diferentes variáveis.

[Gráfico livro de códigos - ver openintro]

Outras variáveis frequentes neste tipo de estudos são o género, a idade, e lugar de nascimento

[Gráfico base de dados]

3. Análise Exploratória: Estatística Descritiva, Multivariada e Explicativa

O raciocínio estatístico parte de um modelo básico que pode ser representado graficamente da seguinte forma:

X -> D
Z

onde
- ***X*** é a variável independente (explanans, explicativas/preditora, causa, input, exógena),
- ***D*** é a variável dependente (explandum, de resposta, efeito, output, endógena), e
- ***Z*** é uma variável de controlo.

Esta representação expressa que as variações de X afectam as variações de D, sem que Z interfira significativamente no resultado. A estatística aborda esta configuração analítica por meio detrês grandes abordagens: estatística descritiva, multivariada e explicativa.

A ***estatística descritiva*** concentra-se na análise individual das variáveis, caracterizando a distribuição dos valores observados a traves do cálculo de estatísticas como a média e a sua

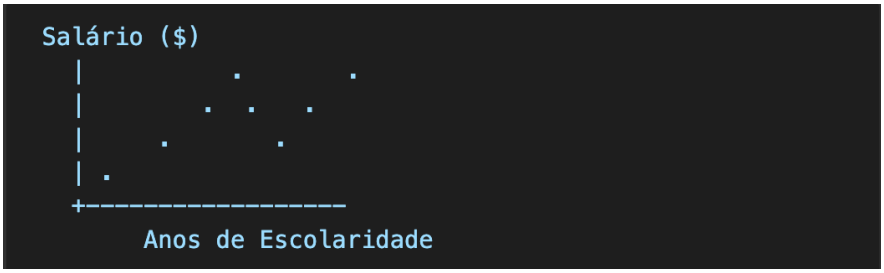
dispersão entre mínimos e máximos. Basicamente, a descritiva fornece ferramentas para organizar e sumariar grandes quantidades de dados em informação com sentido, possibilitando a identificação

Salário (em $)	
Min	1.500
Q1	2.000
Média	2.800
Q3	3.500
Max	5.000

de padrões gerais e desvios particulares. É também chave para identificar e corrigir erros nos dados e introduzir ajustes que permitam uma melhor manipulação dados.

A seguir, é comum que múltiplas variáveis estejam relacionadas entre sim, fazendo parte de uma variável agregada de nível mais geral. Isto é frequente, por exemplo, em certos estudos de percepção que costumam propor um conjunto de perguntas relacionadas com um mesmo assunto (valores medioambientais, satisfação com o funcionamento das instituições, prejuízos raciais). A denominada ***análise multivariada***, com técnicas como analise *factorial*, *de componentes principais* ou de *clusters* se encarrega de detectar e tratar estas associações intrínsecas. Por exemplo, a analise fatorial permitiria que, os dados fornecido por um questionário perguntando sobre Confiança no governo, Confiança no parlamento, Confiança na justiça seja reudizido a um único a um único "indicador de satisfação política", enquanto que um **Análise de Clusters**: agrupa indivíduos com perfis semelhantes.

Finalmente, a ***estatística explicativa*** foca na correlação entre X e D, estimando o impacto da variação na primeira na variação da segunda. Por exemplo, a relação entre anos de escolaridade (X) e salário (D), Trata-se do estudo de um particular tipo de associações, as de tipo de causal. Uma variável está associada ou correlacionada com outra quando as variações dos seus respectivos valores mostram alguma conexão entre si (por exemplo, quando uma se incrementa a outra também).



Entretanto, a estatística é enfática na indicação de que associação não implica causalidade, pelo menos por três razões:

i) A primeira é que pode tratar-se de associações intrínsecas, como as que vimos que estuda a estatística multivariada. Duas variáveis podem estar correlacionadas por serem expressões de uma mesma característica mais geral. A extensão dos braças está correlacionada com a extensão das pernas, mas uma não é a causa da outra, e ambas estão afetadas pela presença de uma terceira como pode ser a altura.

ii) Segundo, a associação entre duas variáveis pode ser exógena mas igualmente mas porem serem explicadas por uma terceira variável não considerada, a nomeada variável de confusão (*'confounding variavel'),* fazendo do vínculo das duas primeiras uma 'relação espúria'. O exemplo clássico é a relação positiva entre o consumo de gelado e o afogamento no mar, duas variáveis de características diferentes e fortemente associadas, mas causadas simultaneamente por uma terceira (período estivo).

iii) Uma terceira razão, de tipo mais epistemologico, se sustenta na impossibilidade de garantir com certeza a presença a presença de causalidade a partir de estudos observacionais, tarefa que só pode ser realizada com desenhos de pesquisa experimentais (OpenIntro:13). Os ***estudos experimentai***s caracterizam-se pela maior capacidade de intervenção dos investigadores sobre a ocorrência do fenómeno estudado, como na construção de grupos de controlo/tratamento aleatórios ou na exposição deles à variável de interesse. Um exemplo claro é a situação de testeo de uma droga (X), que se administra no grupo de tratamento, verificando logo a presença do efeito esperado (Y), e controlado com o grupo de controlo que só recebe placebo (Z). Pelo contrario, os ***estudos observacionais*** se limitam a registrar fenómenos existentes a través de ferramentas com os inquéritos ou recolha de dados secundários sem possibilidade de intervenção voluntária. Estes podem ser ***retrospetivos***, quando olham para eventos já acontecidos, ou ***prospectivos***, quando fazem um seguimento em tempo real do fenómeno estudado. Para a estatística, os estudos observacionais só podem mostrar associações. Não estando controlada pelo investigador a administração de X,a presença da relação causal só pode ser suspeitada (OpenIntro:13).

Falamos de ***estatística exploratória*** quando a analise limita-se a examinar à amostra. Nesta fase, os resultados obtidos só aplicam aos indivíduos observados, não pudendo ser generalizados para a população de estudo. Esta última é tarefa da análise inferencial.

4. Análise Inferencial: os testes estatísticos

Um dos grandes aportes da estatística é a possibilidade de fazer generalização ou inferências de modo rigoroso e contrastável. A analise inferencial completa o ciclo da análise estatística permitindo postular, a partir dos dados observados na amostra, proposições sobre a população não observada.

A inferência estatística pode ser descritiva ou explicativa (King e Verba 1999). De modo bem esquemático, o seu objectivo consiste na aplicação de **testes estatísticos** à estatística exploratória. É aqui onde entram em jogo a celebre 'significancia estatística', com ferramentas como os intervalos de confiança, os margens de erro, os valores p, e as estrelas dos outputs das regressões, assim também como requerimentos vinculados as características da amostra e as propriedades das distribuições de valores das variáveis.

A estatística fornece uma multiplicidade de testes, que variam de acordo com o tipo e número de variáveis envolvidas e o número de observações. Por exemplo, um análise na sua fase exploratória poderia estar indicando que:

> *"65% dos jovens **da amostra** votam a partidos de esquerda"*

A análise inferencial aplicaria os testes correspondentes e, se tudo dar certo, permitiria postular afirmações como que

> *"65% dos jovens **do Portugal** votam a partidos de esquerda,*
> *com um margem de erro de 3% e um 95% de certeza"*

[Figura Panoramica]

5. Tarefas

Seção PANORAMA

1) Identificar e escolher 5 artigos vinculados a um tópico do seu interesse, que utilizem análise estatística.
2) Para cada um dos 5 artigos escolhidos, identificar que tipo de análise aplicam (descritivo, explicativo) e que técnicas e ferramentas (coeficinetes, regressões, boxplots, etc.).
3) Para 2 dos 5 artigos escolhidos,
   a) identificar e descrever a base de dados utilizados (unidade de observação, número de observações, variáveis) e
   b) reconstruir o codebook incluindo pelo menos 5 variáveis.

Seção UMA QUANTITATIVA

4) Para 1 dos 5 artigos escolhidos, identificar a descrição de uma variável numérica. Transcrever essa descrição.

Seção UMA QUALITATIVA

5) Para 1 dos 5 artigos escolhidos, identificar a descrição de uma variável qualitativa. Transcrever essa descrição.

Seção ANALISE BIVARIADA

6) .....

6. Leituras recomendadas

- Capítulo 1 OpenIntro
- Capítulo Bryman

7. Referências

Bryman?

Gerring

OpenIntro

King & Verba

[Intro to EDA –OLI:7]
https://oli.cmu.edu/jcourse/workbook/activity/page?context=90d50f4a80020ca601343b8eb604eef1