DESMITIFICANDO A ESTATÍSTICA

Antônio Fernando de Castro Alves Beraldo*

INTRODUÇÃO

A **Estatística** é um conjunto de métodos que, utilizando procedimentos matemáticos, visa **conhecer** e **descrever** a **realidade** que nos cerca, **analisar** seus fenômenos naturais e sociais e **fornecer informações** de apoio às Ciências. A Estatística é uma das Matemáticas, assim como a Geometria, a Trigonometria e o Cálculo Integral. A Estatística consiste em **contar**, **medir**, **classificar**, **relacionar**, **comparar**, **prever**, **testar** e **analisar** os dados que expressam as características desta realidade.

A palavra "Estatística" foi criada no século XVII, por Helenus Politanus. Seria uma variação do termo "Estadística", algo como "conhecimento das coisas do Estado", "Estado", aqui querendo dizer "governo". A Estatística, em seus primórdios, já foi chamada de "Matemática Política", por Sir William Petty (1623-1687), um dos primeiros economistas, mas a palavra retorna na tradução para a língua inglesa, por John Arthubnot (1667-1735), de um trabalho de Huygens (1629-1695) sobre o cálculo de probabilidades, e em seus trabalhos posteriores.

Os verbos da estatística

Estatística é trabalho de pesquisa, é investigação sobre a realidade. Compreendemos a Estatística através de sua história, uma história em que homens e mulheres se debruçaram sobre os problemas que surgiram, e ainda surgem, desta ação. Ação expressa por verbos, como os seguintes:

Contar

A Estatística começou quando o homem precisou enumerar seus bens, seus rebanhos, as colheitas e os exércitos. Estas informações sempre foram vitais para os reis e governantes, e as contagens eram feitas periodicamente, a seu mando. Temos notícia de levantamentos feitos na China, 2.000 anos A.C., na Bíblia (informações sobre o contingente de soldados e armas do povo judeu), dos recenseamentos no Império Romano (população e extensão territorial)[1], das coletas de dados feitas pelos árabes no século VIII. Ainda na Idade Média, Carlos Magno, rei dos francos e Imperador do Ocidente, organizou o Estado a partir da contagem de seus súditos e de suas propriedades. Guilherme, o Conquistador, ordenou a elaboração do *Domesday Book*, um catálogo dos bens do reino (Inglaterra, 1085) e de sua produção, para fins de ... cálculo e coleta de impostos. Como se pode ver, a Estatística sempre foi associada ao **Estado**. E passou a embasar todas as teorias e práticas da Economia. **Contar** é o método matemático mais primitivo para se conhecer a realidade, mas, por outro lado, talvez seja a maior proeza mental do *homo sapiens*: quantificar algo através de uma abstração – o número.

Medir

Alguns fenômenos não são "contáveis", são "mensuráveis". As técnicas de medir, cuja história se confunde com a história da Ciência, foram amplamente desenvolvidas pela Estatística. A Estatística tem a sua maneira própria de medir, e suas próprias medidas, chamadas **estatísticas**: média, moda, desvio padrão, variância, números-índices ... A Estatística mede, por exemplo, a dispersão (ou concentração) de elementos de um conjunto em torno de um elemento central; a probabilidade da ocorrência de defeitos em um produto industrial; a relação entre o nível de renda de uma população e seu consumo de alimentos; a evolução das taxas de mortalidade de indivíduos acometidos de doenças; a posição de um elétron em torno do núcleo do átomo; a classificação provável de determinado candidato num concurso vestibular com milhares de candidatos; o efeito da propaganda nas vendas de um um produto; a audiência de um programa de televisão; a intenção de votos em um candidato ... Estes processos vêm desde o século XVII, com os estudos de Estatística Demográfica, de John Graunt, e o Cálculo Atuarial, de Edmond Halley. Hoje são técnicas bastante sofisticadas, com farto uso de computadores e muita matemática.

* Prof. do Departamento de Estatística. ICE – UFJF e Assessor de Estatística do CBR – UFJF.

Classificar	**Classificar** é quase uma decorrência natural dos processos de contar e de medir. Medidas estatísticas conduzem à colocação dos fenômenos em **classes**. Classificar pode ser entendido como **categorizar** (colocar em categorias – A, B, C, D ...) ou **ordenar** (colocar em postos: 1º lugar, 2º lugar, 3º lugar, etc.). A Estatística possui também suas medidas especiais de classificação, como as separatrizes e os escores padronizados, entre outras.
Relacionar	A Estatística estuda os relacionamentos entre os fenômenos, no tempo e no espaço. Através de um conjunto de medidas estatísticas, procura-se determinar se existe uma **correlação** (ou interdependência) entre duas ou mais variáveis, e se esta relação, caso exista, é "forte" ou "fraca". Pode-se analisar, por exemplo, a relação existente entre a escolaridade de uma população e a incidência de uma determinada doença; a correlação entre o número de animais predadores em um lugar e os tipos de presas existentes nesta região; o rendimento escolar de alunos e seu quociente de inteligência; o número de acidentes de trânsito e a quantidade de veículos em circulação.
Comparar	Comparar grandezas é uma das áreas onde mais se aplicam os processos estatísticos. São as estatísticas chamadas de **números-índices**, entre outras, de larga utilização na Economia, nas Ciências Sociais, na Medicina, na Administração Pública. Ao comparar valores destas grandezas entre diversos países ou regiões, em épocas diferentes, procura-se, desta forma, medir a evolução destas grandezas – o que fornece os parâmetros para o planejamento governamental das políticas sociais e econômicas, por exemplo.
Prever	As técnicas de previsão estatística, baseadas no Cálculo de Probabilidades, constituem o ferramental básico dos Sistemas de Apoio às Decisões. Principalmente a Análise de Séries Temporais (onde os fenômenos se relacionam diretamente com o passar do tempo), que talvez seja o ramo da Estatística de maior desenvolvimento nos últimos anos. A previsão estatística, conjugada com as técnicas de correlação e de comparação, auxilia no planejamento das ações e no desenvolvimento das empresas, das instituições governamentais e de tecnologia. Uma parte importante da previsão estatística é a **Atuária**, ou Cálculo Atuarial, que trata dos cálculos de seguros (de vida, de acidentes, de doenças, etc.), tendo por base o Cálculo de Probabilidades.
Testar	Testes Estatísticos são processos de verificação da igualdade ou desigualdade entre duas ou mais medidas – entre valores esperados (ou previstos) e valores ocorridos, por exemplo, ou entre estatísticas de dois ou mais conjuntos, separados no tempo ou no espaço. Pode-se testar estatisticamente a eficiência de um processo (uma dieta, por exemplo), ou a eficácia de uma ação (um método de aprendizagem), as diferenças entre os resultados de dois ou mais tipos de tratamentos médicos (a cura pela sugestão, pela alopatia ou pela homeopatia).
Analisar	A Análise Exploratória de Dados é uma evolução nas técnicas de análise de dados. Sem descartar as análises baseadas no Cálculo de Probabilidades, propõe uma visão mais abrangente e profunda da realidade, com ênfase em representações gráficas e estudo de casos nas áreas de Exatas e Saúde, principalmente.

O Método Estatístico

Como foi dito anteriormente, os processos de contar, medir e classificar conjuntos, com um número muito grande de elementos ("conjuntos de tamanho infinito"), se tornaram complexos e trabalhosos, à medida que estes conjuntos cresciam mais e mais de tamanho. Uma coisa é uma costureira tomar as medidas de sua clientela, por mais numerosa que seja; outra coisa é uma indústria de calças *jeans* saber qual deve ser o seu *mix* de produção (entre calças de tamanho P, M, G) de acordo com as medidas de milhões de prováveis consumidores. Um problema é

um pequeno empresário controlar seu fluxo de caixa e sua produção, outra coisa é uma montadora de automóveis prever o seu faturamento e a quantidade de peças de reposição que deve colocar no mercado. Resumindo: nossa questão é conhecer a "realidade", quando esta realidade é complexa, variada, irregular, incerta, e, freqüentemente, mutável. Temos, principalmente, três processos para conhecer a "realidade": o **censo**, o **levantamento** e o **método estatístico**.

Conceitos e termos técnicos da estatística:

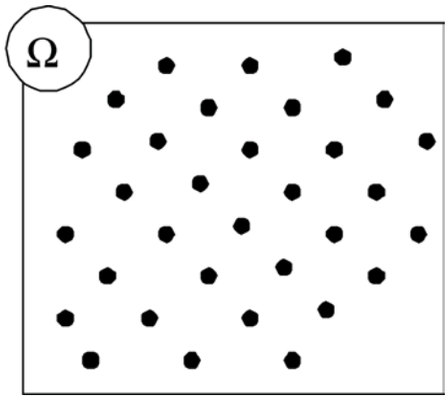

FIG. 1- O Conjunto Universo

Ao lado, na Figura 1 a ilustração do que é entendido como **conjunto universo**, ou **universo**, ou ainda, **população**. Trata-se de uma porção da realidade que nos cerca, sobre a qual é feito nosso estudo. Nosso objetivo é **conhecer** este conjunto, extrair dele informações que nos permitam tomar decisões, comprovar ou rejeitar hipóteses, verificar a evolução de fenômenos, etc. Este conjunto possui as seguintes características:

a) O conjunto universo, notado por Ω, por definição possui um tamanho N suposto infinito ($N \rightarrow \infty$), ou seja, é constituído de incontáveis elementos.

b) Os elementos de um conjunto universo possuem **pelo menos um atributo em comum**.

Exemplos de **conjuntos-universo**:

1 Os alunos da Universidade	Os elementos que pertencem a este conjunto podem ser naturais da cidade ou não, podem pertencer às classes sócio-econômicas A ou B ou C, podem possuir carro ou não, podem ser solteiros, casados, viúvos ou divorciados – mas, para pertencer a este conjunto, têm que estar matriculados em algum curso da universidade, neste momento. Este é o **atributo comum** a todos os elementos do conjunto. Note que, apesar de sabermos exatamente o número de alunos da Universidade, ou seja, podermos contá-los, "este universo é tratado como se seu tamanho fosse infinito".
2 Os alunos da Universidade, do sexo feminino	Este também é um conjunto universo. Agora, seus elementos possuem **dois** atributos comuns: são matriculados na Universidade, e são do sexo feminino.
3 As peças fabricadas na Indústria ABC, no mês de abril, saídas da linha de montagem 3.	Este é um conjunto universo cujos elementos possuem **três** atributos em comum: são fabricados pela indústria ABC, no mês de abril e saíram da linha de montagem 3.

Uma vez definido e delimitado o conjunto universo, podemos estudá-lo a partir de seus **atributos**.

Por exemplo, para o conjunto universo 1 (os alunos da Universidade), podemos analisá-lo sobre os seguintes aspectos (ou atributos): a estatura dos alunos, seu peso, o curso em que estão matriculados, se são naturais da cidade ou não, seu índice de rendimento acadêmico, sua classe sócio-econômica, o meio de transporte utilizado para chegar ao *campus*, sua intenção de voto nas eleições, e muitos outros.

Todo este conjunto de dados é processado, fornecendo informações que descrevem a realidade que nos cerca. A coleta de dados é feita, principalmente, por três métodos: o **Censo**, o **Levantamento**, e pelo **Método Estatístico**.

Censo: ou recenseamento é o processo de coleta de dados em que todo o conjunto universo é pesquisado. Todos os elementos do conjunto são estudados, uma a um, e o censo só termina quando todo o conjunto universo for totalmente abrangido. Censos são trabalhosos, demorados, custam caro e, por isso, são realizados por órgãos do governo. Censos demográficos são realizados de cinco em 5 anos, quando uma grande quantidade de recenseadores é recrutada para coletar dados sobre a população, através de questionários. Desta forma, podemos medir a evolução de dados como a população das cidades e do meio rural, as taxas de natalidade e mortalidade, as características de raça, o credo religioso, as migrações internas, enfim, dados demográficos, sócio-econômicos, culturais, de escolaridade e da saúde da população.

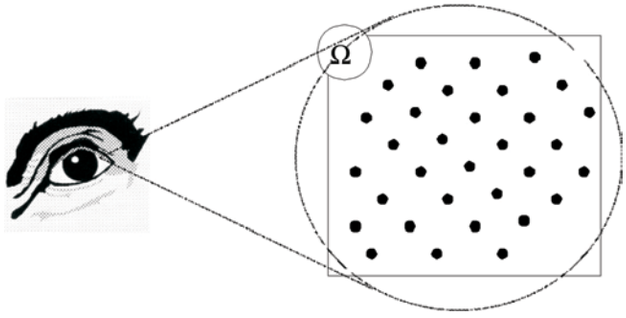

Figura 2: O conjunto Universo (Ω)

Para se ter uma idéia da magnitude do problema do Censo em um país como o nosso, com seus 8,5 milhões de km2, e 170 milhões de habitantes:

- *O Brasil conta com 5.507 municípios. As populações destes municípios variam desde Borá, SP, com 789 habitantes (397 homens e 392 mulheres - 50,3 % de homens e 49,7 % de mulheres) até São Paulo, SP, a maior cidade do país, com seus estimados 11 milhões de habitantes.*
- *No Censo 2000, foram recenseados cerca de 42 milhões de domicílios. Foram impressos mais de 100 milhões de questionários, 30.000 mapas de cidades, vilas e localidades e mais de 200 mil croquis (desenho do setor) para orientar os recenseadores na sua área de trabalho.*
- *Para fazer o censo demográfico, o IBGE contratou cerca de 151 mil recenseadores. Os recenseadores foram contratados após concurso realizado em todo o país, e ainda foram contratados mais técnicos e analistas, chegando a 200 mil pessoas envolvidas no trabalho.*

O Censo começou em agosto de 2000, em quase todos os municípios do país. Ocorreram alguns problemas no nordeste, devido às enchentes, mas, transcorrido um mês, já tinham sido coletados dados sobre 44 milhões de habitantes - o cronograma estava em dia. Os resultados preliminares do Censo começam a ser divulgados em dezembro de 2000. Prevê-se para meados de 2001 o início da publicação dos resultados consolidados, que deve terminar em 2003.

No Brasil, contagens primitivas foram feitas desde os séculos XVII e XVIII, utilizando-se registros paroquiais, e estimativas feitas pelos Ouvidores da Corte portuguesa. Em 1750, a Coroa Portuguesa ordenou um censo sobre brasileiros aptos para a defesa do território.

Em 1852 tentou-se realizar um censo, mas a população revoltou-se, imaginando que o censo fosse uma tentativa de capturar negros fugidos ou alforriados. Somente em 1872 foi feito um primeiro censo nacional organizado, que se repetiu de 10 em 10 anos até 1940 (não foram feitos censos em 1910 e 1930). Em 1934 foi fundado o IBGE, Instituto Brasileiro de Geografia e Estatística, que tem organizado e realizado os censos, antes decenais, agora feitos de 5 em 5 anos, juntamente com a amostragem domiciliar.

Nos EUA, no final do século XIX, foi feita uma licitação para a realização do censo demográfico de 1890. A concorrência foi vencida por Hermann Hollerith, que inventou uma série de máquinas e dispositivos que poderiam realizar a coleta e processamento dos dados. Adaptou máquinas de escrever e de perfurar cartões, criou um sistema de codificação de dados em cartões de papelão e conseguiu entregar os resultados antes do tempo previsto. Este trabalho foi a origem a empresa IBM, uma das gigantes da computação mundial. E foi a origem também da palavra "holerite", que significa contracheque de pagamento de salário.

Levantamentos: são parecidos com o censo, mas são realizados em um subconjunto do universo, chamado **partição**, "escolhido" segundo informações anteriores que indicam que aquele subconjunto é bastante "representativo" do universo.

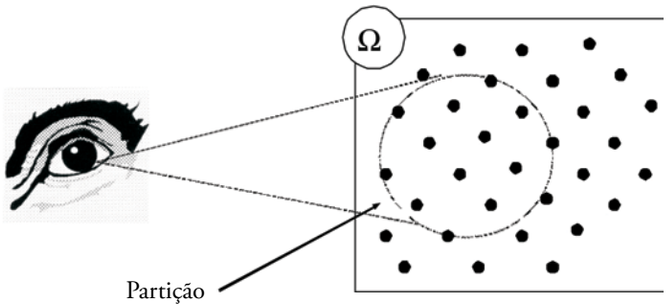

Figura 3: Universo e Partição

Levantamentos são menos onerosos do que os censos, e são muito utilizados para verificar as diferenças entre as partes de um mesmo universo. Pode-se dizer que um levantamento é um censo feito sobre uma das partes do universo.

O **Método Estatístico** surgiu a partir das dificuldades em se fazerem censos e levantamentos. Aplicando este método, podemos descrever universos de qualquer tamanho, analisar as relações entre seus elementos e efetuar todos os processos estatísticos de previsão e testes. Consiste nas seguintes etapas:

1 O Universo em estudo é definido e delimitado, e alguns de seus elementos são "sorteados" para compor um subconjunto do universo, chamado **Amostra**. Antes de efetuar este sorteio, o universo é **homogeneizado**, isto é, deve-se garantir que "cada elemento do Universo tem a mesma probabilidade de ser sorteado que qualquer outro elemento".

2 A amostra, assim obtida, é de tamanho muito menor do que o Universo, sendo, portanto, passível de ser estudada, isto é: podemos calcular, sobre a amostra, uma série de medidas que **descrevem** a amostra. Estas medidas descritivas da amostra são chamadas de **estatísticas**. São medidas como a média e a mediana, a variância e o desvio padrão, os quartis, e outras.

As **estatísticas** descrevem a amostra, ou melhor, **traduzem em números** a constituição e a relação entre seus elementos.

3 A partir das estatísticas (que descrevem a amostra), são efetuados uma série de cálculos matemáticos com o objetivo de determinar outros números, que são chamados de **parâmetros**. Estes parâmetros são medidas estatísticas que descrevem o Universo. O cálculo dos parâmetros é chamado de **Inferência Estatística**

Atributos e Variáveis

Anteriormente, falamos de **atributos**, que seriam algo como as qualidades ou características que os elementos de um Universo (e das amostras dele extraídas) possuiriam. Estes atributos podem ser valorados, isto é, assumir diferentes valores, **numéricos ou não**, passando a se chamar **variáveis.**

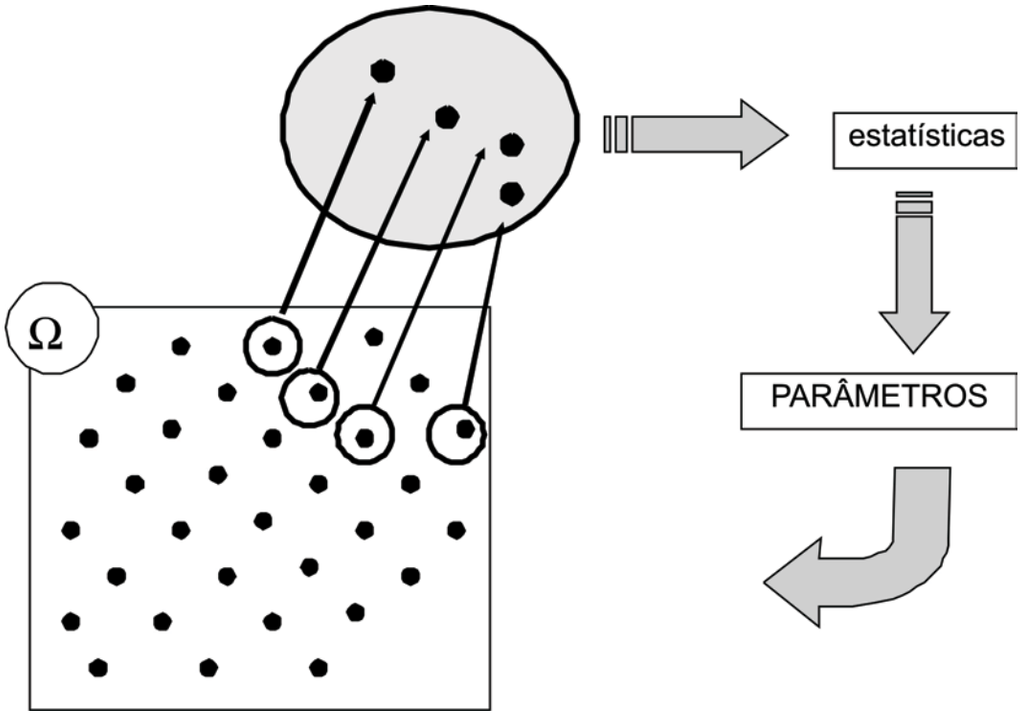

Figura 4: O Ciclo do Método Estatístico

Nota: embora os processos de amostragem e inferência sejam baseados no Cálculo de Probabilidades, muitas amostras são coletadas por outros processos, não-probabilísticos ou intencionais. Mesmo assim, continuam válidas, sob determinadas condições, as etapas descritas acima.

No exemplo dado, a variável **estatura** dos estudantes pode assumir qualquer valor entre, digamos, 1,20 m e 2,10 m. Já a variável **classe sócio-econômica** pode assumir os valores convencionais A, B, C, D ou E. A variável **naturalidade** (se o elemento é natural da cidade) pode assumir os valores S ou N (sim ou não).

As variáveis são classificadas de acordo com o **tipo de valores** que podem assumir. Temos, portanto, em uma classificação inicial, os seguintes tipos de variáveis:

VARIÁVEIS ALEATÓRIAS	QUANTITATIVAS **(Numéricas)**	DISCRETAS
		CONTÍNUAS
	QUALITATIVAS (Categóricas ou Nominais)	

Uma variável é dita **quantitativa** ou **numérica** quando assume exclusivamente valores numéricos. É quantitativa **discreta** quando estes valores pertencem ao conjunto dos Naturais, mais o zero: 0, 1, 2 , 3 , 4 , 5 ... Geralmente, estes valores são resultado de um processo de **contagem**. Uma variável é quantitativa (ou numérica) **contínua** quando pode assumir valores pertencentes ao conjunto dos Racionais[1]. Geralmente, estes valores são resultado de uma **medição**.

Uma variável é dita **qualitativa** ou **categórica** ou **nominal** quando pode assumir apenas valores não-numéricos. No exemplo que estudamos, o **curso** em que o aluno está matriculado é uma variável qualitativa, que pode assumir os valores "Engenharia", "Medicina", "Direito", "Enfermagem", etc. Um caso especial das variáveis qualitativas ou categóricas é o das **variáveis lógicas**, que podem assumir apenas dois valores: S e N (sim ou não).

Comentários

1. *Discretas ou Contínuas? Notar que as variáveis numéricas discretas podem ser **tratadas como se fossem contínuas**. Um dos problemas que são resolvidos pela Estatística, como foi dito, é o de efetuar contagens em conjuntos muito grandes. Mesmo para estes conjuntos (o número de analfabetos no país, por exemplo), o resultado desta contagem pode ser perfeitamente calculado, com uma precisão razoável, usando um método estatístico chamado **Estimação**. Sem entrar em detalhes, neste momento, podemos dizer que **estimar** uma quantidade é calcular um **intervalo numérico** em que o **valor mais provável** de uma medida esteja nele contido. Note que escrevemos "intervalo numérico", ou seja, a grosso modo, "entre dois números". Diz-se que uma pessoa tem entre 240.000 a 340.000 fios de cabelo, isto é, ela tem **entre** 240 mil **e** 340 mil fios de cabelo. Este resultado é obtido assim: divide-se a área total do couro cabeludo do cidadão em quadradinhos de área igual, digamos, 1 cm^2 de área. Para simplificar, vamos supor que o couro cabeludo contenha 1.000 quadradinhos. Sorteia-se uma série de quadradinhos, digamos, uns trinta quadradinhos. Em cada quadradinho sorteado conta-se o número de fios de cabelo, e calcula-se a **média** de "fios de cabelo por quadradinho". Calcula-se também uma outra estatística, chamada **desvio padrão**, que é, por assim dizer, a "faixa de variação" da média. Se a média foi de 290 fios de cabelo por quadradinho, e o desvio padrão de 50 fios de cabelo por quadradinho, dizemos que o número de "fios de cabelo, por quadradinho", está entre 240 e 340. Como são 1.000 quadradinhos, dizemos que a pessoa possui entre 240.000 e 340.000 fios de cabelo. Note que "número de fios de cabelo" é, a priori, uma variável numérica discreta. Quando seu valor se torna muito grande, dá-se a ela um **tratamento** de variável numérica contínua.*

2. *Variáveis Categóricas Lógicas: este tipo de variável também é muito utilizado pela Estatística. Dissemos que ela pode assumir os valores S e N (sim e não). Estendendo o raciocínio, podemos dizer que esta variável pode assumir dois valores, opostos e complementares, e que são **mutuamente excludentes**, ou seja: a variável possui dois estados, que não podem ocorrer simultaneamente. Por exemplo: "cara" ou "coroa", no lançamento de uma moeda; "masculino" ou "feminino", no nascimento de uma criança; "ligado" ou "desligado", para um aparelho elétrico.*

3. Por outro lado, podemos substituir as categorias de uma variável qualitativa por números,

se esta variável qualitativa possui um caráter hierárquico ou ordinal, ou mesmo de graduação em nível ou intensidade. Por exemplo, em uma pesquisa de opinião pública a respeito do presidente da república, as respostas possíveis são: "ótimo", "bom", "regular", "ruim" ou "péssimo" (variáveis qualitativas). Devido ao alto grau de subjetividade nesta conceituação, pode-se substituir a pergunta da pesquisa por outra: "Qual nota, numa escala de 0 a 10, você daria ao Presidente da República?". Com este procedimento, tenta-se tornar a pesquisa mais objetiva, com a utilização de variáveis quantitativas. O inverso pode também ser utilizado: as famílias de um bairro podem ter uma classificação sócio-econômica A, B, C, D ou E (variável qualitativa) conforme sua renda familiar (variável quantitativa).

4. *Muitas vezes você encontrará variáveis qualitativas, **identificadas** por números. Por exemplo, em um questionário acerca do estado civil de um elemento amostral, pode-se identificar "solteiro" por "01", "casado" por "02", "divorciado" por "03 etc. É preciso não confundir estes valores, digamos, pseudonuméricos, com valores de uma variável quantitativa. Estado civil é uma variável qualitativa e deve ter o tratamento correspondente, adequado. Outro exemplo: no seu número de matrícula, consta, digamos, o dígito "04" - que corresponde ao curso no qual você está matriculado. Apesar de ser um número, estes dígitos representam variáveis qualitativas.*

5. *Existem outras classificações de variáveis, quanto ao tipo. Por exemplo, um tipo de dados chamados **ordinais**. Referem-se ao lugar, ou **posto**, que ocupam em uma coleção disposta em ordem. Esta ordem pode ser 1º , 2º , 3º , 4º lugar ... ou "muito frio", "frio", "morno", "quente", "muito quente"... Notar que estes dados, apesar de serem escritos também de forma numérica, não são valores de uma variável quantitativa, por não estarem sujeitos a muitas das operações matemáticas.*

6. *Outra classificação de variáveis muito encontrada atualmente é a de **dados de razão**. São valores de uma variável que estão referenciados a uma base. Por exemplo, valores de temperaturas (podemos usar as escalas Celsius, Kelvin ou Farenheit – cada uma delas possui valores diferentes para uma mesma temperatura, dependendo da base, ou do "zero" da escala, atribuído a um estado de uma substância – por exemplo, om ponto do gelo da água). Geralmente as variáveis utilizadas na Física são classificadas como dados de razão: pressão, intensidade luminosa, velocidade, e assim por diante.*

Em nosso dia-a-dia, utilizamos variáveis para identificar e mesmo caracterizar completamente os componentes da "realidade" em torno de nós. Por exemplo figura da página 84.

NOTAS

[1] Os habitantes do Império Romano tinham que responder ao *census* na sua cidade de origem, e a punição para quem fugisse ou não respondesse era a pena de morte. Segundo a Bíblia, os pais de Jesus, Maria e José, tiveram que empreender uma viagem de Nazareth, na Galiléia, para Belém, na Judéia, para responder ao Censo ordenado por Cesar.

[2] De uma maneira mais rigorosa e abrangente, qualquer valor pertencente ao conjunto dos números reais.

Variáveis qualitativas:
- Raça
- Cor e aspecto do pelo
- Imperfeições e traços distintivos
- Origem
- Linhagem

Variáveis quantitativas
- Peso
- Tamanho
- Idade
- Número de Filhotes (ninhada)
- Quantidade de ração ingerida

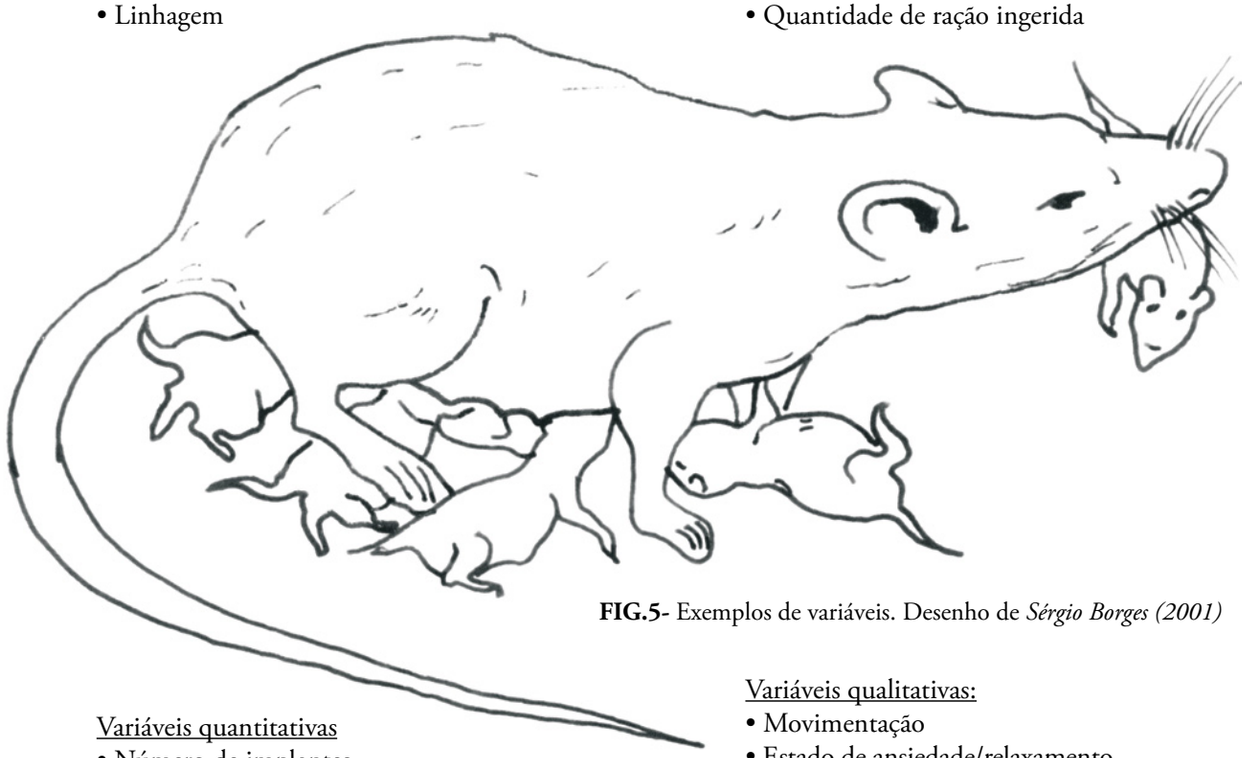

FIG.5- Exemplos de variáveis. Desenho de *Sérgio Borges (2001)*

Variáveis quantitativas
- Número de implantes
- Número de fetos vivos e mortos

Variáveis qualitativas:
- Movimentação
- Estado de ansiedade/relaxamento
- Agressividade
- Grupo (controle, tratamento)