Pós-Graduação em Economia e Gestão em Saúde

Módulo de Estatística Aplicada

Prof.ª Dr.ª Maria Dolores Montoya Diaz

ÍNDICE

1. Conceitos Básicos

1.1 Fases de uma Análise Estatística

a) Definição do Objetivo: é a etapa principal, pois a partir da idéia clara do alvo a ser atingido é que será traçada a estratégia de trabalho. É nesta fase que se deve avaliar se o objetivo pretendido é factível, e quais as alternativas existentes também gerariam resultados úteis e interessantes. Qualquer equívoco nesta etapa certamente gerará frustrações na fase de análise dos resultados.

b) Definição das Informações Necessárias - Definição da População e da Amostra : nesta fase se determinam quais as informações devem ser coletadas a fim de se atingir o objetivo pretendido. Isto implica em detalhar, também, o procedimento a ser utilizado na coleta, que pode ser por exemplo, pesquisa de mercado ou um levantamento em bancos de dados primários tais como, o Sistema Integrado de Acompanhamento Financeiro do Município , do Estado ou da União. Esta etapa exige a definição da abrangência da população sob estudo, bem como da dimensão da amostra a ser avaliada.

➔ ***Lembrete**: População é o conjunto de todos os elementos aos quais estão associadas determinadas características que se gostaria de identificar, conhecer ou mensurar.*

Amostra consiste em uma parte dos elementos da População.

▪ **Exemplo 1**: objetivo = conhecer a estatura média da população brasileira.

população = toda a população brasileira

amostra = 1.000 brasileiros escolhidos aleatoriamente

▪ **Exercício 1**: objetivo = conhecer o percentual médio de gastos com pessoal da Prefeitura de São Paulo.

população = ?

amostra = ?

c) Levantamento das Informações Disponíveis : esta fase consiste na efetiva pesquisa e registro ordenado dos dados.

d) Análise Exploratória dos Dados - Estatítica Descritiva: nesta etapa, as informações serão exploradas com dois objetivos básicos: identificar eventuais erros de coleta - crítica - ou de registro e identificar a presença de alguns fenômenos ou de relações entre os elementos que estão sendo estudados.

e) Análise dos Resultados = Conclusões: neste estágio serão realizados testes para confirmar ou rejeitar algumas hipóteses levantadas inicialmente.

➔ **IMPORTANTE:** Não se esqueça, porém, de que qualquer análise estatística somente gerará bons resultados quando combinar o conhecimento acurado das técnicas com bom senso !!!

2. Análise Exploratória dos Dados

2.1 Tabelas e Gráficos

De acordo com o que foi destacado anteriormente os dados coletados devem ser registrados e organizados de modo a que se possa obter as informações mais diretamente sem desperdício de recursos e de tempo.

➔ **Cuidado**: a forma de entrada dos dados não precisa necessariamente a forma mais adequada para analisá-los. Todos os softwares de planilha eletrônica possuem funções

que permitem reordenar as informações de maneira rápida e fácil. Assim, antes de iniciar o trabalho de digitação dos dados, compare sempre o formato com que as informações estão disponíveis e a forma mais fácil de se trabalhar com elas. Verifique se é possível, por intermédio de recursos do software a ser utilizado, reordenar as informações. Caso contrário, avalie qual a estratégia será menos custosa em termos de digitação e processamento das informações. Em suma, **PLANEJE** o seu banco de dados.

O Microsoft Excel, por exemplo, fornece um conjunto de comandos para facilitar o gerenciamento de uma lista ou banco de dados. No Menu principal, dentro da opção DADOS, selecione CLASSIFICAR para dispor as linhas de acordo com a ordem de uma determinada coluna. Outras possibilidades estão disponíveis, tais como FILTRO (localiza e identifica um subconjunto de seus dados a partir de determinadas características especificadas como critérios para a seleção das informações), TABELAS DINÂMICAS (sintetiza e permite alguns tipos de análise sobre os dados de um banco de dados). Para maiores informações, consulte o manual do software ou a ajuda interativa.

A construção de gráficos também se constituem em um ótimo instrumento de trabalho pela facilidade de visualização e compreensão dos resultados gerados. No Excel, existe um Auxiliar Gráfico, acionado por um Botão que  permite a elaboração rápida de um gráfico.

Existe, entretanto, um tipo de gráfico que desempenha um papel particularmente importante em qualquer análise estatítica que é o HISTOGRAMA ou Gráfico de Distribuição de Freqüência.

Um exemplo bem simples pode ilustrar bem a utilidade deste instrumento. Pretende-se examinar o padrão de composição de algumas famílias em termos de números de filhos. Os dados disponíveis são os seguintes:

Quadro 1

Família	Número de Filhos
A	1
B	2
C	3
D	2
E	3
F	3
G	3
H	4
I	4
J	4
K	5
L	6
Total de Famílias	12

A partir daí, deve-se proceder da seguinte maneira:

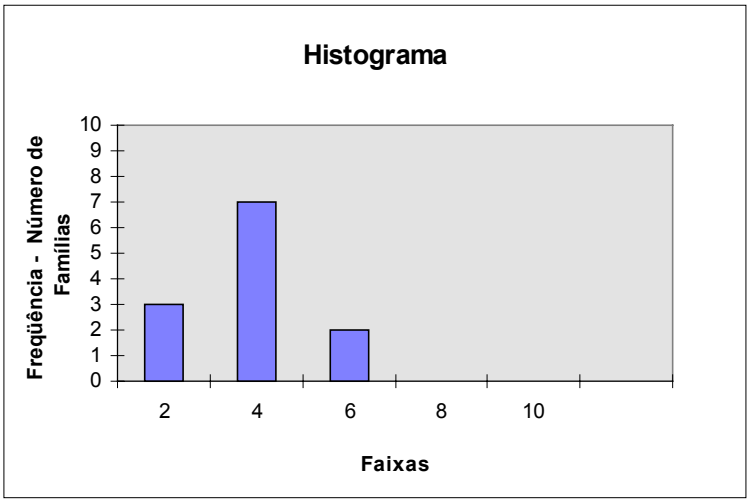
i. organizar os dados em uma tabela, que estabeleça faixas e o número de ocorrências em cada faixa, conforme se verifica abaixo. Esta tabela recebe o nome de DISTRIBUIÇÃO DE FREQÜÊNCIAS:

Tabela 1

Faixas	Freqüência
até 2 filhos	3
mais de 2 até 4 filhos	7
mais de 4 até 6 filhos	2
mais de 6 até 8 filhos	0
mais de 8 até 10 filhos	0
Mais de 10 filhos	0
Total	12 famílias

ii. da forma como as informações foram rearranjadas é possível construir um Histograma, que consiste em um gráfico de barras em que a altura das barras corresponde à freqüência com que os valores representados pelas faixas apareceram. No caso, a altura das barras corresponde ao número de famílias que possuiam o o número de filhos representados pelas faixas expressas no eixo das variáveis X. O gráfico permite a visualização imediata de que na amostra colhida existe a predominância de famílias com mais de 2 até 4 filhos.

Gráfico 1

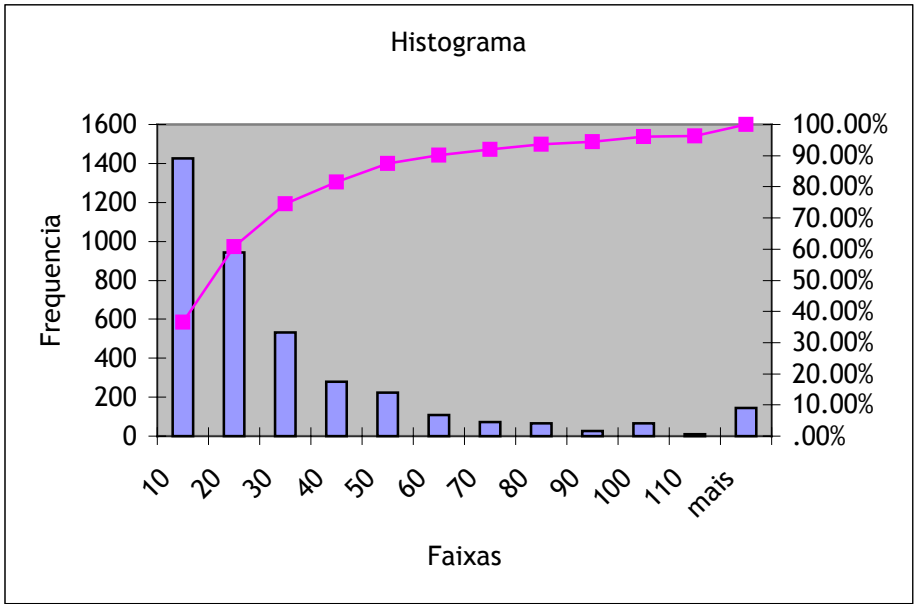


Considerando, agora, uma situação mais realista em que se queira analisar a distribuição dos gastos mensais com medicamentos de um conjunto de 3897 indivíduos, deve-se proceder da mesma maneira. Assim, deve-se construir a tabela com a DISTRIBUIÇÃO DE FREQÜÊNCIAS. Os resultados podem ser vistos abaixo:

Faixas de Gastos Mensais(R$)	*Freqüência - casos observados*	*% cumulativo*
10	1427	36.62%
20	944	60.84%
30	531	74.47%
40	278	81.60%
50	225	87.37%
60	108	90.15%
70	71	91.97%
80	67	93.69%
90	27	94.38%
100	67	96.10%
110	9	96.33%
mais	143	100.00%
Total	3897	

De forma análoga à apresentada anteriormente, será construído o Histograma, o qual pode ser visualizado a seguir:

Gráfico 2



Nota-se facilmente que existe uma predominância nas faixas iniciais, ou seja, até gastos de R$ 60,00 mensais. Assim, tem-se uma visualização do perfil sintético dos gastos realizados com medicamento. Para verificar se compreendeu, experimente construir um histograma com uma desagregação maior das faixas. Compare com o perfil mostrado acima.

➔ **OBS**: No Excel, a elaboração da Tabela de Distribuição de Freqüência, bem como do Histograma correspondente é automática a partir de funções específicas agrupadas

dentro da opção **ANÁLISE DE DADOS** que é acessada a partir do menu **FERRAMENTAS**. Se a alternativa ANALISAR DADOS não estiver disponível será necessário instalá-la – na mesma opção FERRAMENTAS (TOOLS), optar por SUPLEMENTOS (ADD-INS) e selecionar os itens FERRAMENTAS DE ANÁLISE(ANALYSIS TOOLPAK). Para maiores detalhes verificar no manual ou no HELP interativo do software.

O MÓDULO ANALISAR DADOS DO MENU UTILITÁRIOS SERÁ ESSENCIAL PARA ESTE CURSO[1].

2.2 Estatística Descritiva

Na seção anterior, procurou-se mostrar formas alternativas de organizar informações estatísticas, de tal forma que ficassem evidenciados alguns aspectos interessantes. Pode-se, de outro modo, calcular algumas medidas que resumem algumas das caracterísicas presentes na série de dados.

Todas as medidas apresentadas a seguir, podem ser calculadas a partir da opção ESTATÍSTICA DESCRITIVA, dentro do módulo ANALISAR DADOS, com exceção da Covariância e do Coeficiente de Correlação, que apresentam-se como itens separados dentro do módulo ANALISAR DADOS.

As funções estatísticas do EXCEL também permitem o cálculo de todas as medidas consideradas. A escolha entre um ou outro método dependerá da preferência do usuário.

[1] Todos os procedimentos estatísticos podem ser elaborados manualmente sem a utilização desta ferramenta. Porém, a montagem deste conjunto de procedimentos é extremamente trabalhosa. Assim, em decorrência da disponibilidade da procedimentos automáticos existentes na opção ANALISAR DADOS optou-se pela utilização desta facilidade.

2.2.1 Medidas de Tendência Central

Representam a série de dados pelos seus valores médios, identificando assim, a posição da distribuição dos valores sobre o eixo X.

2.2.1.1 Média Simples e Ponderada

A média aritmética simples pode ser representada pela seguinte fórmula:

ARITMÉTICA SIMPLES: $$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} = \sum_{i=1}^{n} X_i \times \frac{1}{n} = \sum_{i=1}^{n} X_i \times \text{peso do item i}$$

que nada mais representa do que a soma de todos os elementos (n ao todo) de uma amostra, dividida pelo número total de elementos da amostra, ou seja, n.

É importante destacar que quando se fala em média simples, existe uma pressuposição de que todos os elementos envolvidos possuem a mesma importância. Caso isto não ocorra, é necessário incorporar uma ponderação para estes elementos. Assim, fórmula para a média aritmética ficaria:

ARITMÉTICA PONDERADA: $\bar{x} = \sum_{i=1}^{n} X_i \times \text{peso do item i}$, que é diferente para cada um dos elementos

É preciso ter em mente que a média tem uma interpretação geométrica, ou seja, no caso da situação representada noGráfico 1, a média de filhos por família é de 3,33. Assim, este valor pode ser colocado dentro da segunda faixa, de tal forma que os dados se posicionassem ao seu redor de forma equilibrada.

2.2.1.2 Mediana

Se os valores da série forem ordenados em ordem crescente, a mediana é o elemento que ocupa a posição central.

Se o número de elementos, n, for ímpar, a mediana será o elemento central, ou seja, será o elemento de ordem $\frac{n+1}{2}$. Assim , se a série tiver 7 elementos, a mediana será o 4º elemento ($\frac{7+1}{2}$).

Se o número de elementos, n, for par, a mediana será a média entre os elementos de ordem $\frac{n}{2}$ e $\frac{\mathrm{n}}{2}+1$. Assim, no exemplo simples do Quadro 1, em que estão sendo considerados 12 elementos, a mediana corresponde à média entre o 6º (n/2) e o 7º (n/2+1) elementos. Como no caso, ambos tem o valor de 3, a mediana assume o valor 3[2].

2.2.1.3 Moda

A Moda corresponde ao valor mais freqüente da distribuição. No caso do exemplo anterior, o valor da Moda também será 3. Deve-se destacar, entretanto, que esse valor é igual à Mediana apenas POR COINCIDÊNCIA.

2.2.2 Medidas de Dispersão

Estas medidas tem como função verificar a representatividade dos resultados fornecidos pelas medidas de tendência central. Assim, se a dispersão dos valores considerados for grande as medidas de tendência central serão menos representativas.

---

[2] Se você não está conseguindo chegar a esse valor a partir do Quadro 1, não se esqueça que é preciso ORDENAR os dados de forma crescente.

Pode-se, por exemplo, ter duas séries com as mesmas médias que podem ter um histograma completamente diferente em decorrência de apresentarem valores muito dispersos.

2.2.2.1 Amplitude

Corresponde à diferença entre o maior e o menor valor da série de dados,ou seja, à distância entre os valores extremos. No caso do nosso exemplo, a amplitude é de 5 (6 filhos foi o maior valor observado e 1, o menor).

2.2.2.2 Variância

Para medir a dispersão dos valores da série em relação à média, pode-se pensar que basta simplesmente somar os desvios em relação à média. Porém, se fizer esta experiência verificará que não foi uma boa idéia, mas faça essa experiência com os dados do exemplo dos filhos.

Assim, será necessário considerar os desvios independentemente do sinal da diferença. E isso será feito por meio do cálculo da diferença ao quadrado. Tem-se, então:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \Rightarrow \text{variância populacional}$$

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \Rightarrow \text{variância amostral}$$

O incoveniente destes indicadores é que representam valores cuja unidade de medida corresponde ao quadrado da unidade de medida original. Se, por exemplo, estão sendo considerados valores correspondentes a preços de uma determinada mercadoria, a

unidade de medida da variância será $(R\$)^2$, cuja utilidade prática não é significativa. Por esse motivo, adquire relevância o Desvio-Padrão, apresentado a seguir.

2.2.2.3 Desvio-Padrão

A grande utilidade do Desvio-padrão está justamente no fato de apresentar a mesma unidade de medida da série original, permitindo uma conclusão imediata acerca da maior ou menor dispersão dos valores da série. As fórmulas a serem utilizadas são as seguintes:

$\sigma = \sqrt{\sigma^2}$ ⇨ desvio-padrão populacional

$S = \sqrt{S^2}$ ⇨ desvio-padrão amostral

2.2.2.4 Coeficiente de Variação

Conforme destacado acima, as medidas de dispersão mostradas estão associadas à unidade de medida da série de dados. Com isso, a tarefa de comparação das dispersões de valores de séries distintas fica bem comprometida. Para transpor esta barreira, pode ser utilizado o Coeficiente de Variação que corresponde a:

$$Coef.Var. = \frac{\sigma}{\bar{x}}$$

Quanto menor o valor do Coeficiente de variação maior será a representatividade da média.

2.2.3 Relações entre Variáveis

Até esta altura, apresentaram-se medidas para analisar uma série, independentemente de qualquer influência de outros fatores. Ocorre, entretanto, na

maioria dos casos, pretende-se verificar também a existência de alguma relação com elementos de outras séries de dados.

2.2.3.1 Covariância

A covariância se assemelha muito à variância, como o próprio nome sugere. A única diferença reside no fato de que aqui se trata de análise da dispersão de duas séries, ou seja, pretende-se avaliar se o padrão das discrepâncias em relação à média ocorrem de forma similar, oposta ou nada tem em comum. A visualização da fórmula de cálculo da covariância permite entender mais claramente essa idéia:

$$Cov(x,y) = \sigma_{x,y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \times (y_i - \bar{y})}{n-1}$$

Compare com a fórmula da variância amostral, e verifique se y=x, a covariância entre x e y corresponde à *variância* de x.

De forma análoga ao que se verifica na variância, a unidade de medida da covariância entre x e y, corresponde ao produto das unidades de medida de x e y. Sendo assim, uma outra fórmula será necessária para eliminar o inconveniente gerado por este fato.

2.2.3.2 Coeficiente de Correlação

O Coeficiente de Correlação desempenha papel análogo ao do Coeficiente de Variação dentro do contexto de análise conjunta de duas séries. Assim, tem-se que:

$$Coef.Correl. = \frac{Cov(x,y)}{\sigma_x \times \sigma_y}$$

Se o Coeficiente de Correlação for exatamente igual a 1, pode-se dizer que existe uma correlação positiva perfeita entre as duas séries. Assim, ambas apresentam o mesmo

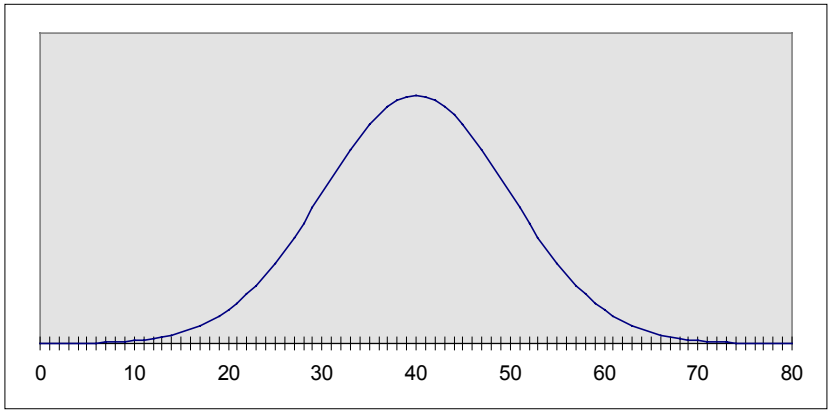
comportamento, ou seja, quando uma aumenta, por exemplo, a outra acompanha. Se o Coeficiente de Correlação for exatamente igual a 0, não se pode identificar nenhuma relação entre as variáveis. No caso do Coef. de Correlação ser igual a -1, existe uma correlação negativa perfeita, ou seja, quando uma série vai em uma direção a outra segue na direção oposta.

Obviamente, o cálculo do Coeficiente de Correlação não gera necessariamente algum dos três valores citados acima, apesar de obrigatoriamente estar contido no intervalo entre -1 e 1. Valores próximos a cada um dos três sugerem conclusões semelhantes às apresentadas, com a diferença de que no caso da existência de correlação positiva ou negativa, esta não será perfeita.

3. Distribuições de Probabilidade

3.1 Normal

A mais famosa das distribuições de freqüência é a Distribuição Normal:



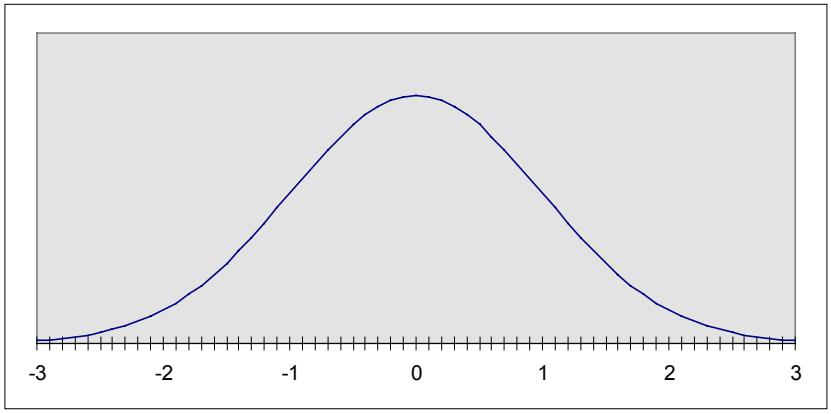
Distribuição Normal com média=40 e desvio-padrão=10

A grande vantagem desta distribuição está no fato de que a partir de apenas dois parâmetros - a média e o desvio-padrão - é possível calcular a probabilidade associada

a qualquer valor que se queira. Assim sendo, cada par de valores distintos de média e desvio-padrão gera uma curva diferente, o que obviamente também geraria dificuldades. Normalmente, para representar esta distribuição utiliza-se a seguinte notação: N(média-μ,desvio padrão-σ).

Justamente para contornar esta complicação, trabalha-se com a Normal Padronizada, cuja média é 0 e a variância é igual a 1.

A partir de uma transformação simples de variável, é possível chegar a uma variável que tenha essa distribuição normal:



Distribuição Normal Padronizada ⇨ tem média=0 e desvio-padrão=1

Conforme destacado anteriormente, o objetivo principal, QUE NÃO DEVE SER ESQUECIDO, é de associar probabilidades à ocorrência de algum fenômeno sob observação. Assim, por exemplo[3], sabe-se que a distribuição dos salários anuais de auxiliares de escritório de uma grande empresa é Normal com média igual a $ 12.500 e desvio padrão igual a $2.800 e se quer calcular três indicadores:

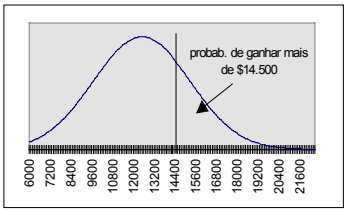
a) proporção dos auxiliares de escritório que ganham mais de $14.500

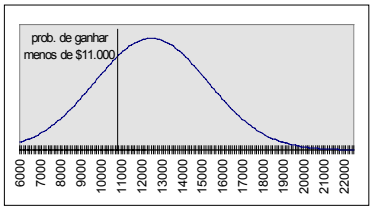
[3] Este exercício está em Lapponi(1997), pág.200

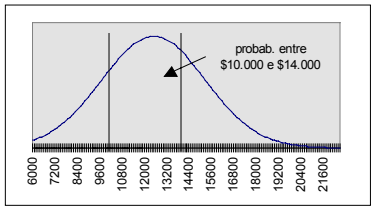
b) proporção dos auxiliares de escritório que ganham menos que $ 11.000

c) proporção dos auxiliares de escritório que ganham entre $10.000 e $14.000

Graficamente a idéia torna-se bem simples. Então, tem-se que:







Estas probabilidades podem ser encontradas por meio da utilização das tabela da Dsitribuição Normal Padronizada em Anexo, ou diretamente pela aplicação da função estatística Dist.Norm() do EXCEL. Obviamente, a aplicação da função gera resultados mais precisos do que a tabela em decorrência da inexistêcia de limitações em termos de casas decimais utilizadas. Os resultados são os seguintes:

a) 23,75% - função e 23,89% pela tabela; b)29,61% pela função e 29,81% pela tabela; c) 51,80% pela função e 51,52% pela tabela.

Obviamente, nem todos os fenômenos a serem estudados seriam bem representados por uma Distribuição Normal. Assim, foram estudadas várias outras distribuiçòes de probabilidade, das quais serão apresentadas três, em decorrência de sua maior aplicação.

3.2 Qui-quadrado

A soma de variáveis Normais Padronizadas elevadas ao quadrado apresenta esta distribuição. Diferentemente, da Distribuição Normal que necessitava de dois parâmetros para o cálculo de probabilidades, neste caso, têm apenas um parâmentro que é o número de graus de liberdade. Normalmente, esta distribuição é representada

por χ^2 (v), onde v corresponde aos graus de liberdade. Esta distribuição terá papel relevante na realização de testes de hipótese, da mesma forma que as duas distribuições apresentadas a seguir.

3.3 t

Sendo Z uma variável aleatória com distribuição N(0,1), e Y uma variável aleatória com distribuição χ^2 (v), então se for construída uma variável T, que resulta da divisão de Z por Y, esta terá uma distribuição denominada *t* , cujo único parâmetro será v, ou seja, o número de graus de liberdade.

3.4 F

Sendo Y_1 uma variável aleatória com distribuição χ^2 (v_1), e Y_2 uma variável aleatória com distribuição χ^2 (v_2), então se for construída uma variável F, que resulta da divisão de Y_1 por Y_2, esta terá uma distribuição denominada *F* , cujos parâmetros serão v_1 e v_2, ou seja, os números dos graus de liberdade de Y_1 e Y_2, respectivamente.

4. Testes de Hipótese

Hipótese estatística é uma suposição sobre o valor de algum parâmetro populacional. A comprovação ou rejeição dessa suposição pode ser obtida por meio da realização de um teste.

Quando se realiza um teste de hipótese, na verdade, acaba-se lidando com duas hipóteses, necessariamente excludentes, ou seja, se uma se verifica a outra não pode ocorrer. São elas:

- H_0 ou Hipótese nula, que deve ser construída de tal forma a ser rejeitada;
- H_a ou Hipótese alternativa, corresponde àquela suposição que se quer comprovar.

A esta altura, é preciso lembrar que estamos lidando com incerteza, portanto, a aceitação ou rejeição de uma hipótese estará sempre condicionada à existência de algum erro. Assim, convencionou-se definir dois tipos de erro:

- Erro tipo 1 que consiste em Rejeitar H_0 quando ela é verdadeira;
- Erro tipo 2 que consiste em Aceitar H_0 quando ela é falsa.

Sendo assim, uma boa estratégia será minimizar o erro tipo 1 pois rejeitando-se H_0 ter-se-á uma maior "certeza" acerca da conclusão adotada. Percebe-se, desse modo, que o erro tipo 1 desempenha um papel importante na análise, e a ele está associada uma probabilidade de ocorrência, conhecida como Nível de Significância do teste. Um nível de significância pequeno, por exemplo, de 5% implica que, se H_0 for rejeitada existe 5% de chance de se estar cometendo um erro.

Os procedimentos básicos para realização de um teste são os seguintes:

i. definir H_0 e H_a
ii. fixar o nível de significância
iii. a partir dos dados amostrais calcular o valor do parâmetro sobre o qual estão definidas H_0 e H_a
iv. comparar com os limites determinados pelas tabelas da distribuição correspondente
v. concluir pela aceitação ou não da H_0

4.1 Diferença de Médias

Freqüentemente, o analista se depara com a necessidade de realizar comparações entre amostras distintas. Assim, pode-se citar a situação em que se pretende verificar se as médias salariais de determinada categoria são distintas em cada Estado, ou se custos de determinado tipo apresentam médias distintas nas várias filiais de uma mesma empresa, ou se dois tipos de ativos geram médias semelhantes de rendimento,

entre outras. Na verdade, o que se pretende é verificar se existe diferença entre as médias de duas populações distintas.

Para realizar este tipo de comparação basta efetuar um teste de diferenças entre médias.

As premissas para a realização do teste, como bem colocou Lapponi(1997) são as seguintes:

i. "As duas populações têm médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2 ;
ii. São extraídas duas amostras independentes, uma de cada população, de tamanhos n_1 e n_2 e médias, $\overline{x_1}$ e $\overline{x_2}$.
iii. A diferença das duas médias $\overline{x_1}$ - $\overline{x_2}$ é uma nova variável aleatória, que será maior que 0 (zero) quando $\overline{x_1} > \overline{x_2}$, e menor que 0 (zero) quando $\overline{x_1} < \overline{x_2}$ ".
iv. A variância dessa nova variável é $\overline{x_1}$ - $\overline{x_2}$ é igual à soma das variâncias σ_1^2 e σ_2^2 , ou seja, $\sigma^2_{\overline{x}_1-\overline{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
v. $H_0 = \mu_1 - \mu_2 = 0$

$H_a = \mu_1 - \mu_2 \neq 0$

Na verdade, pode-se considerar a existência de algumas situações aprticulares que são as seguintes:

1. As variâncias populacionais σ_1^2 e σ_2^2 são conhecidas - uso de distr. normal;
2. As variâncias populacionais σ_1^2 e σ_2^2 são desconhecidas, porém assume-se que são iguais - uso da distribuição t ;
3. As variâncias populacionais σ_1^2 e σ_2^2 são desconhecidas e assume-se que são distintas - uso de distribuição t, com correção.

Um exemplo será de grande valia para compreender os conceitos envolvidos. Suponha que estejam disponíveis as seguintes séries:

Série 1	Série 2
20	25
23	26
26	25
22	24
24	27
	25

Observe que as séries apresentam uma quantidade distinta de elementos. No caso de se considerar a possibilidade 1, no EXCEL, a escolha das opção *Teste-z: duas amostras para médias* dentro do menu ANALISAR DADOS, gerará os seguintes resultados:

Teste-z: duas amostras para médias		
	Série 1	*Série 2*
Média	23	25.3333333
Variância conhecida	5	1.06666667
Observações	5	6
Hipótese da diferença de média	0	
z	-2.15003291	
P(Z<=z) uni-caudal	0.01577625	
z crítico uni-caudal	1.644853	

P(Z<=z) bi-caudal	0.00788812	
z crítico bi-caudal	1.95996108	

Como o módulo de z calculado é maior, tanto quando se realiza o teste uni, como bi-caudal, pode-se concluir pela não aceitação de H_0, ou seja, pela existência de diferenças de média entre a Série 1 e a Série 2. Outra forma de chegar a esta conclusão seria pela comparação do valor de P(Z<=z) uni-caudal ou bi-caudal com o valor de 5% para o nível de significância escolhido. Quando esse valor for menor que o nível de significância, rejeita-se H_0.

Se for considerada a possibilidade 2, ou seja a opção *Teste-t: duas amostras presumindo variâncias equivalentes*, os resultados serão os seguintes:

Teste-t: duas amostras presumindo variâncias equivalentes		
	Série 1	Série 2
Média	23	25.3333333
Variância	5	1.06666667
Observações	5	6
Variância agrupada	2.81481481	
Hipótese da diferença de média	0	
gl	9	
Stat t	-2.29676286	
P(T<=t) uni-caudal	0.02362449	

t crítico uni-caudal	1.83311386	
P(T<=t) bi-caudal	0.04724899	
t crítico bi-caudal	2.26215889	

Neste caso, também rejeita-se H_0, pois ambas as probabilidades são menores de 5%.

Finalmente, se for considerada a possibilidade 3, ou seja a opção *Teste-t: duas amostras presumindo variâncias diferentes*, os resultados serão os seguintes:

Teste-t: duas amostras presumindo variâncias diferentes		
	Série 1	Série 2
Média	23	25.3333333
Variância	5	1.06666667
Observações	5	6
Hipótese da diferença de média	0	
gl	5	
Stat t	-2.15003291	
P(T<=t) uni-caudal	0.04212076	
t crítico uni-caudal	2.015049176	
P(T<=t) bi-caudal	0.08424152	
t crítico bi-caudal	2.570577635	

Observe que o valor da estatística calculada é o mesmo que se verifica no caso 1, ou seja, -2.15003291. A única diferença entre as situações é a de que estes esta variável não tem distribuição t. Sendo assim, deve ser introduzida uma correção nos valores da tabela da distribuição t[4]. Em termos de resultados, se for adotado o teste uni-caudal conclui-se pela rejeição de H_0 e se for bi-caudal, a conclusão será opsta, ou seja, pela aceitação da hipótese nula de igualdade de médias.

4.2 Diferença de Variância

Este tipo de teste tem aplicação semelhante ao de diferença de médias. A distinção reside na utilização da Distribuição F. No exemplo acima, os resultados obtidos pela utilização da opção seriam os seguintes:

Teste-F: duas amostras para variâncias		
	Série 1	Série 2
Média	23	25.3333333
Variância	5	1.06666667
Observações	5	6
gl	4	5
F	4.6875	
P(F<=f) uni-caudal	0.0604729	
F crítico uni-caudal	5.19216314	

Ao nível de significância de 5%, aceita-se a hipótese nula de igualdade de variâncias entre as duas séries, pois P(F<=f) uni-caudal é maior que o nível de significância. À mesma conclusão se chegaria, analisando o valor de F calculado que é de 4.68, inferior, portanto, ao nível crítico de 5.19.

Qual a implicação deste resultado sobre os testes de diferença de médias realizados na seção anterior?

[4] Maiores detlhes podem ser encontrados em Lapponi(1997) e Hoffman.

5. Regressão Linear

É um método que procura descrever e analisar a relação entre uma determinada variável, Y, cujo comportamento se quer explicar, e uma ou mais variáveis, contidas em uma matriz denominada X que se supõem servirem como explicativas da evolução de Y. Assim, tem-se que:

$$Y = f(X)$$

Pode-se citar como exemplo, um modelo em que se pretende analisar a evolução do Consumo Agregado. A teoria econômica indica que uma boa variável explicativa neste caso seria a Renda. Neste caso, tem-se que:

$$Consumo = f(Renda)$$

Ocorre, por outro lado, que as relações econômicas não são tão exatas assim, pois inúmeros fatores acabam influenciando, de maneira indireta o comportamento da variável Y, tais como a aleatoriedade do comportamento humano. Há ainda a considerar a possibilidade de existirem elementos que não são incorporados por desconhecimento do analista quanto à sua importância, ou pela dificuldade de mensuração deste fator.

Por esse motivo, nos modelos de regressão além das variáveis explicativas torna-se necessária a inclusão de um termo aleatório, u, responsável pela incorporação de todos esses elementos.

Desse modo:

$$Y = f(X, u)$$

Nos modelos de regressão assume-se que a relação entre a variável explicada, Y, e as explicativas contidas na matriz X é linear, ou ao menos linearizável, no caso de relações não-lineares. Tem-se, então:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + u$$

Esta estrutura associada a um conjunto de hipóteses tanto sobre o termo aleatório como sobre a matriz X, permite que, a partir da estimação dos parâmetros β da equação acima que se estabeleça a relação existente entre a variável explicada e as explicativas.

O método de estimação convencionalmente utilizado é o Método dos Mínimos Quadrados, que em essência busca minimizar a soma dos quadrados dos resíduos, ou seja, minimizar a diferença entre os valores observados e os estimados a partir da equação da regressão.

Obviamente, as considerações apresentadas acima permitem apenas uma visão superficial dos conceitos, aplicações e restrições envolvidas na construção de um modelo de regressão. Maiores aprofundamentos devem ser buscados em Kmenta, Pindyck e Rubinfeld, Wonnacott e Wonnacott, entre outros.

Na área de saúde, inúmeras são as aplicações desta categoria de modelos. Um exemplo encontra-se abaixo.

Pretende-se avaliar qual a relação existente entre os gastos com medicamentos e o nível de renda dos indivíduos. Abaixo, encontram-se os resultados obtidos. O coeficiente da variável rendom indica a existência de uma relação positiva entre o comportamento da renda e os gastos com medicamentos. Além disso, verifica-se que, a cada R$ 100,00 de incremento na renda haverá, em média, um incremento de R$0,24 nos gastos mensais com medicamentos.

MODELO 1: Gastos com Medicamentos como função da Renda Domiciliar

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	27.9058	2.4779	11.2621	0.0000%
rendom	0.0024	0.0008	3.0567	0.2253%

Um mapeamento deste tipo de relação permite que ações, tanto de políticas públicas como empresariais, possam ser direcionadas mais especificamente ao grupo-alvo.

Bibliografia

Fonseca, J. S. e Martins, G. A.(1996) - *Curso de Estatística*, Editora Atlas.

Hoffman, R. - *Estatística para Economistas*, São Paulo: Livraria Pioneira Editora.

Kmenta,J. - *Elementos de Econometria*, vol. 1 e 2, Editora Atlas.

Lapponi, J.C. (1997) - *Estatística usando Excel 5 e 7*, São Paulo: Lapponi Treinamento e Editora.

Mittelhammer, R.C.(1996) - Mathematical Statistics for Economics and Business, New York: Springer-Verlag.

Pindyck, R.S. e Rubinfeld, D.L.(1991) - *Econometric Models and Economic Forecasts*, 3rd edition, McGraw-Hill.

Wonnacott, R.J. e Wonnacott,T.H. - *Econometria*, Livros Técnicos e Científicos Editora.