Introdução à Análise de Dados em Ciências Sociais

Nelson do Valle Silva
Carlos Antonio Costa Ribeiro

Conteúdo

PARTE 1: DESCREVENDO VARIÁVEIS, CORRELAÇÕES E ASSOCIAÇÕES 3

CAPÍTULO 1 – MENSURAÇÃO E DESCRIÇÃO DE DADOS 4

CAPÍTULO 2 - ASSOCIAÇÃO: RPE E DESVIO DA INDEPENDÊNCIA 24

CAPÍTULO 3 - ANÁLISE DE VARIÂNCIA 25

CAPÍTULO 4 – ANÁLISE DE REGRESSÃO 33

PARTE 2: PROBABILIDADE E INFERÊNCIA 51

CAPÍTULO 5 – PRINCÍPIOS BÁSICOS DE PROBABILIDADE 52

CAPÍTULO 6 – A DISTRIBUIÇÃO BINOMIAL 61

CAPÍTULO 7 – AMOSTRAGEM E INFERÊNCIA 74

CAPÍTULO 8 – TESTE DE HIPÓTESE (E ESTIMAÇÃO COM VARIÂNCIA DESCONHECIDA) – (3 AULAS: TH1, TH2 E T-STUDENT) 85

PARTE 3: MÉTODOS PARA ANÁLISE DE ASSOCIAÇÃO 100

CAPÍTULO 9 – ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS 101

CAPÍTULO 10 – ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS E QUANTITATIVAS (ANOVA) (REPETE CAPÍTULO 2 ATÉ PÁGINA 4) 112

CAPÍTULO 11 – REGRESSÃO LINEAR SIMPLES E MÚLTIPLA 119

PARTE 1: DESCREVENDO VARIÁVEIS, CORRELAÇÕES E ASSOCIAÇÕES

CAPÍTULO 1 – MENSURAÇÃO E DESCRIÇÃO DE DADOS

I – Mensuração e Descrição de Dados

O que é Estatística? Existe confusão e controvérsia mesmo entre os profissionais da área, quanto à definição e delimitação do campo próprio de interesse desse ramo da matemática. A concepção mais tradicionalmente aceita vê a estatística definida a partir de duas grandes funções. A primeira dessas funções é a "descrição de dados", consistindo basicamente na sumarização de dados de forma a torná-los mais facilmente utilizáveis. A segunda função é a chamada "indução estatística", ou seja, o estabelecimento de princípios que permitam generalizações sobre uma população a partir do exame de uma amostra tirada dessa população. A descrição e a indução caracterizam dessa forma as duas grandes áreas de definição da Estatística. Mas, como veremos ao longo deste curso, não podemos separar totalmente essas duas áreas, a utilização de estatísticas descritivas sendo fundamental ao processo de inferência. Antes de entrarmos em maiores considerações a respeito da Estatística Descritiva e da Estatística Indutiva nos será útil o estabelecimento de alguns conceitos preliminares.

I.1 - Alguns Conceitos Preliminares

Consideremos primeiramente o conceito de unidade de observação: qualquer objeto ou evento cujas características são suscetíveis de serem empiricamente observáveis. Unidade de observação é, portanto, qualquer segmento da realidade do qual observações sistemáticas podem ser feitas a respeito de suas características. De fato, normalmente não estamos interessados nas unidades de observação em si, mas fundamentalmente na variação de uma dada característica ou no relacionamento entre características ao longo das unidades de observação. Quando tratarmos das variações de uma dada característica entre as diversas unidades de observação nos referimos a elas como variáveis: variável é uma característica das unidades de observação cujas diferenças de estado podem ser detectadas através da aplicação de algum esquema classificatório pré-definido.

Ao conjunto total de unidades cujas características estamos interessados em observar chamamos população. Exemplos de populações são "todos os habitantes de uma área geográfica bem definida"; "todos os alunos em uma sala de aula"; "os parafusos produzidos por uma fábrica ou máquina" etc. Normalmente a população de interesse engloba um número excessivamente grande de unidades de observação, de tal forma que o estudo dessa população não é viável. Por isso, obtemos uma amostra dessa população, ou seja, uma coleção de unidades de observação selecionadas de tal forma que são garantidamente "representativas" daquela bem definida população.

O sentido do conceito de "representatividade" de uma amostra ficará claro ao longo do curso, uma vez que nosso interesse primordial é a Inferência Estatística, o estudo dos

princípios pelos quais uma amostra "representativa" pode ser obtida e a partir da qual generalizações sobre a população podem ser feitas.

O emprego de um esquema classificatório para detectar diferenças de estado em uma variável é chamado de mensuração. Assim, sobre as unidades de observação selecionadas na amostra aplica-se um esquema em que números ou símbolos são designados para a especificação das diferenças de estado de uma ou mais variáveis. Podemos representar a relação dos elementos (unidades de observação) de uma amostra juntamente com os resultados das mensurações obtidas para suas diversas variáveis de interesse através de uma matriz de dados, com o seguinte aspecto geral

	Variáveis (*j*)			
elementos (*i*)	1	2	...	J
1	v_{11}	v_{12}	...	v_{1j}
2	v_{21}	v_{22}	...	v_{2j}
			...	
I	v_{i1}	v_{i2}	...	v_{ij}

onde v_{ij} indica o valor da variável *j* obtido pela aplicação de um dado esquema de mensuração ao elemento (unidade de observação) *i*.

Com base na matriz de dados podemos distinguir duas grandes etapas da investigação social de base estatística: a coleta e a análise de dados. A primeira etapa é definida por Galtung da seguinte forma: "dada uma matriz de dados vazia, preenche-la com valores. A tarefa de coleta de dados termina quando todas as células (ou pelo menos quase todas) tenham sido preenchidas com as respostas ou valores apropriados". Essa etapa envolve, portanto, de um ponto de vista da estatística, o processo de amostragem bem como a mensuração das variáveis de interesse. A segunda etapa, análise de dados, consistirá essencialmente na manipulação da matriz de dados com o objetivo de verificar hipóteses ou estimar valores relativamente à população que deu origem à amostra. A análise de dados corresponde dessa forma à Indução Estatística num sentido restrito.

I.2 – Escalas de Mensuração

Como já vimos, a mensuração consiste na aplicação de um esquema de números ou símbolos para especificar diferenças de estado de uma variável. A esse esquema damos o nome de escala de mensuração. Grosso modo, podemos distinguir três tipos de escalas de mensuração, correspondendo a níveis diferentes de propriedades matemáticas associadas à cada esquema: escalas nominais, escalas ordinais e escalas de intervalo ou razão. A cada tipo de escala teremos um instrumental matemático-estatístico adequado ao tratamento das variáveis mensuradas nesse nível de mensuração, satisfazendo aos requisitos impostos pelas referidas propriedades matemáticas a ela associados. Examinemos em mais detalhe cada um desses tipos de escala.

I.2.1 – Escalas Nominais

A mensuração ao nível nominal, o tipo mais simples de mensuração possível, pode ser entendida como sendo a designação de números ou símbolos na formação de subclasses que representam características únicas. Em outras palavras, consiste na atribuição do elemento observado a uma categoria parte de um conjunto com as seguintes propriedades:

a) as categorias são mutuamente exclusivas – ou seja, nenhum elemento observado pode ser classificável em mais de uma categoria.

b) as categorias são exaustivas – ou seja, nenhum elemento pode deixar de ser classificado. O esquema de classificação deve englobar todos os casos possíveis.

Suponhamos a título de exemplo que uma variável mensurada ao nível nominal seja "tipo de ocupação". Um esquema classificatório possível, que atende ás propriedades enunciadas acima poderia ser

categoria 1: ocupação manual
categoria 2:ocupação não-manual

A operação formal básica do nível nominal de mensuração é a da equivalência, simbolizada matematicamente pelo sinal =. A relação de equivalência é caracterizada por duas propriedades, a simetria e a transitividade:

a. Simetria : se $x=y$ então $y=x$.
b. Transitividade: se $x=y$ e $y=z$ então $x=z$.

É importante que se observe que embora por vezes se associe números às diversas categorias, não faz sentido a execução de qualquer operação assimétrica envolvendo uma variável medida a nível nominal. Por exemplo, o número da carteira de identidade é uma variável nominal, sendo esse número apenas usado para identificar o indivíduo, tanto quando o seu próprio nome. Não faz nenhum sentido se somar os números de identidade, da mesma forma que não faria se estivéssemos somando os nomes dos indivíduos.

I.2.2- Escalas Ordinais

Um nível mais elevado de mensuração é alcançado quando além de podermos classificar as observações em categorias exaustivas e mutuamente exclusivas, podemos também ordenar essas categorias em termos do grau ou intensidade com que essas observações possuem uma dada característica, embora não se possa precisar o quanto dessa caracteriza cada observação possui. Assim, entende-se por mensuração ordinal a designação de símbolos ou números no propósito de se identificar uma relação de ordem entre as diversas observações em termos de alguma característica, essa ordenação não envolvendo nenhum sistema especifico de intervalos. Um exemplo de escala ordinal poderia ser 'nível ocupacional', envolvendo as seguintes ordens

1. Trabalhadores na agricultura
2. Ocupação manuais não-agrícolas
3. Ocupações não-manuais

Observe-se que nesse caso existe uma característica subjacente que tentamos mensurar através da escala (que procura corresponder ao conceito de 'status social'); no entanto, embora possamos ordenar os indivíduos em termos dessa característica usando a escala acima, não podemos precisar a distancia que separa as diversas ordens.

Formalmente, além das propriedades vistas para as escalas nominais, as escalas ordinais possuem como operação básica aquela representada pelo sinal > (maior que). Assim, as propriedades básicas das escalas ordinais são a assimetria e a transitividade:

a. assimetria: se $x>y$, então $y \not> x$
b. transitividade: se $x>y$ e $y>z$, então $x>z$

Similarmente ao caso das variáveis nominais, não faz sentido a execução de operações aritméticas envolvendo variáveis mensuradas a nível ordinal, uma vez que não possuímos informações a respeito da magnitude das distancias entre as diversas ordens.

I-2.3 – Escalas de Intervalo ou de Razão

A mensuração ao nível de intervalo consiste em designar números no propósito de identificar relações de ordem de alguma característica, tendo as ordens intervalos iguais, embora arbitrários. Por exemplo, a renda dos indivíduos numa amostra pode ser mensurada em reais, numa escala de zero ao maior valor possível, com intervalos iguais de uma unidade de real (ou centavos). Um outro exemplo poderia ser o da mensuração de temperatura. Duas escalas são freqüentemente utilizadas na mensuração dessa característica: a escala Celsius (certificada) e a escala Fahrenheit. Ambas são em nível de intervalo, diferindo no entanto quanto à unidade de mensuração e quanto ao ponto de origem (zero) de cada escala. Assim temos graficamente:

Congelamento			Fervura da Água	
0°	33,4°	66,7°	100°	Escala Celsius
+32°	92°	152°	212°	Escala Fahrenheit

Apesar disso, ambas contém a mesma quantidade e espécie de informações, ou seja, dada uma medida numa escala pode-se obter uma medida na outra, através da relação

$$F=95C+32$$

onde F é um numero de grau de escala fahrenheit e C o número de graus centigrados.

Com as escalas de intervalo temos um nível de mensuração em que podemos efetuar operações aritméticas. Podemos mensurar distâncias entra observações e comparar a magnitude dessas distancias.

A escala de intervalo que possui um zero não arbitrário é dita de "razão", constituindo um nível de mensuração um pouco mais elevado do que o nível de intervalo. Exemplificando, suponhamos a mensuração de comprimento. De novo podemos mensurar essa característica segundo duas escalas usualmente utilizadas, a de polegadas e a métrica. Graficamente, uma jarda seria:

0"	12"	24"	36"	polegadas
0"	30,48"	61,96"	91,44"	centímetros

Podemos ver que agora o ponto de origem zero não é mais arbitrário, como no caso das escalas de temperatura, indicando agora um ponto em que a característica medida desaparece.

Com as escalas de razão podemos não só medir e comparar distâncias como podemos comparar diretamente os pontos através do cálculo de razões. Exemplificando, enquanto que no caso da temperatura não podemos dizer que 40ºC é o dobro de 20ºC porque o ponto de origem zero é arbitrário, no caso das escalas de comprimento podemos dizer que 2 metros é o dobro de 1 metro uma vez que a origem não é mais arbitrária.

I.3- Estatística Descritiva

A função básica da estatística descritiva é como já vimos a de sumarização das informações que precisamos a respeito de um dado conjunto de observações. Evidentemente, a forma mais completa de descrição de um conjunto de dados é sua

listagem exaustiva, observação por observação. Mas, também é evidente que quanto maior for o conjunto de observações, mais difícil será a apreensão das propriedades deste conjunto. Como, por exemplo, comparar duas listagens de observações, cada uma consistindo de, digamos, 1000 observações? Obviamente necessitamos de informações sumarizadoras que retenham certas propriedades básicas do conjunto.

Assim, tendo sido preenchida a matriz de dados pelas operações de amostragem e mensuração, a primeira tarefa da etapa de análise de dados será a de examinar o comportamento das características mensuradas, ou seja, descrever e sumarizar as distribuições (o conjunto) das diversas variáveis. Mais uma vez, o conceito fundamental a ser retido é o de "sumário", uma vez que a análise descritiva de uma distribuição nada mais é que a substituição da listagem exaustiva da variável (a distribuição dessa variável) por medidas que descrevam sumariamente essa distribuição. Como variáveis podem ser mensuradas a níveis distintos de mensuração, temos para cada nível medidas descritivas (sumarizadoras das distribuições) adequadas àquele nível especifico. Teremos, pois, medidas adequadas a sumarização de variáveis mensuradas ao nível nominal, outras adequadas ao nível ordinal e, finalmente, outras adequadas ao nível intervalo ou razão.

I.3.1- Medidas Descritivas Adequadas a Atributos (Nível Nominal)

Suponhamos uma característica nominal A, compreendendo duas categorias, a ea. A matriz de dados em que A está listada é composta de 15 elementos, ou seja, levantamos 15 unidades de observação. Suponhamos ainda que a listagem exaustiva da característica A seja

a,a,a,a,a,a,a,a,a,a,a,a,a,a,a

A sumarização permitida pelo nível nominal de mensuração é aquela que envolve apenas a contagem das ocorrências de cada tipo de atributo, ou seja, o levantamento das freqüências de cada categoria da variável analisada. No caso do exemplo envolvendo a variável A acima poderíamos construir a seguinte tabela sumarizadora da listagem dessa variável:

Variável A	
Categoria	Freqüência
a	6
a	9
Total	15

A essa tabela é usualmente dado o nome de "distribuição de freqüências" da variável. Generalizando para qualquer atributo *X* composto de *m* categorias, o aspecto da tabela de distribuição de freqüências será:

Variável X	
Categoria	Freqüência
x1	n1
x2	n2
xm	nm
Total	N

onde xi indica as categorias do atributo X, ni as respectivas freqüências e N é o número total de observações.

Tendo-se obtido as freqüências de cada categoria, pode-se proceder a uma análise comparativa dessas categorias em termos de três medidas sumarizadoras: razões, proporções e porcentagens.

Chama-se razão a uma comparação (divisão) entre duas categorias tendo como base uma das categorias. Conceitualmente a razão representa o numero de elementos na categoria "numerador" para cada elemento da categoria "denominador". Exemplificando, suponhamos a variável "Ocupação" envolvendo as categorias "manual" e "não-manual", com freqüências respectivamente 90 e 10. A razão manual/não-manual (aqui denotada R) seria:

$$R = \frac{\text{manual}}{\text{não-manual}} = \frac{90}{10} = 9$$

O que significa dizer que para cada indivíduo com ocupação não-manual existem 9 indivíduos com ocupação manual.

Para cada razão existirá sempre sua inversa, ou seja, no nosso exemplo podemos também calcular a razão não-manual/manual que é

$$\frac{\text{não-manual}}{\text{manual}} = \frac{10}{90} = 0{,}11 = \frac{1}{R}$$

Como é difícil a interpretação de uma razão decimal, como no caso visto acima é habitual multiplicar-se a razão por uma constante (normalmente 100), obtendo-se então o numero de casos da categoria "numerador" para cada 100 casos da categoria "denominador". Dessa forma, para a razão não-manual/manual, temos 11 indivíduos de ocupação não-manual para cada cem indivíduos da ocupação manual.

Uma proporção é a fração da freqüência de uma dada categoria sobre a freqüência total. No caso do exemplo da variável ocupação a proporção de indivíduos com ocupação manual é

$$p_{manual} = \frac{\text{Frequencia "manual"}}{N} = \frac{90}{90+10} = 0{,}9$$

Quando uma proporção é multiplicada por cem, indicando o número de casos da categoria em um total de cem casos, temos o que se denomina uma percentagem. Para o mesmo exemplo que acima, a percentagem de indivíduos como ocupação manual é

$$\%manual = p_{manual} 100 = 0{,}9100 = 90\%$$

ou seja, ocorrem 90 casos de indivíduos com ocupação manual para cada 100 indivíduos ao todo.

Resumindo mais formalmente

razão: $\frac{n_i}{n_j}$

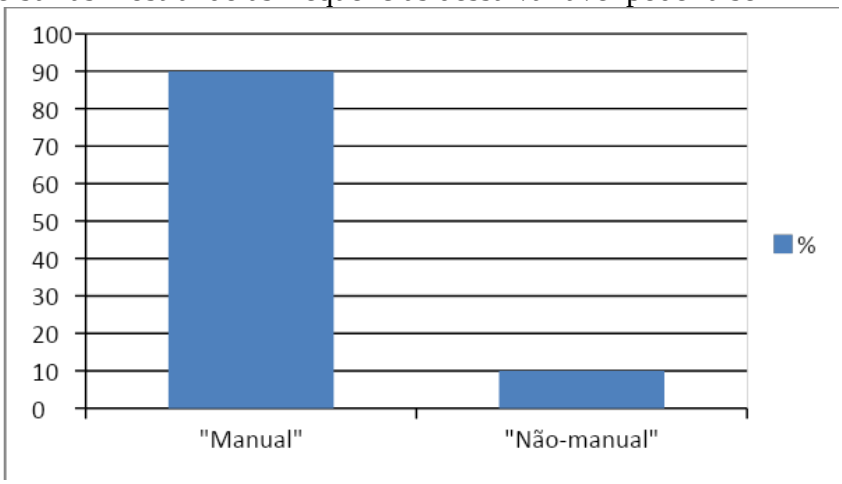
medidas adequadas ao nível nominal

proporção: $n_{ij} = \frac{1}{m_{xj}} = \frac{n_i}{N}$

percentagem: $100x_{ij} = \frac{1}{m_{xj}} = 100x\frac{i}{N}$

Além das medidas mencionadas acima, podemos utilizar no resumo da distribuição de um atributo a camada categoria modal. Entende-se por categoria modal a categoria de maior freqüência. Ela nos informa a categoria 'típica' de um atributo: para o nosso exemplo poderíamos dizer que "tipicamente" os indivíduos em nosso conjunto de observações (população ou amostra) desempenham ocupações manuais, uma vez que essa é a categoria mais freqüente para essa variável.

A representação gráfica usual para variáveis manuais é o diagrama de barras. Basicamente, o diagrama de barras consiste em representar em dois eixos ortogonais as freqüências das categorias, reservando para o eixo horizontal a discriminação das categorias e para o eixo vertical suas freqüências absolutas ou relativas (proporções ou percentagens). Por exemplo, no caso da variável ocupação utilizado anteriormente, um diagrama de barras mostrando as freqüências dessa variável poderia ser



I.3.2- Medidas Adequadas ao Nível Ordinal

Os procedimentos vistos para o nível nominal são também adequados ao nível ordinal. Além disso, outros procedimentos são também apropriados ao nível ordinal de mensuração.

A idéia de ordem é compatível com a de acumulação. Em termos práticos, isto quer dizer que alem de verificarmos as freqüências associadas às diversas ordens, podemos igualmente computar a "freqüência acumulada" ao longo de uma escala ordinal. Uma distribuição de freqüências acumuladas pode ser entendida como um quadro que mostra a freqüência de ocorrência das observações até e inclusive um dado nível da escala ordinal. Por exemplo, suponhamos que as respostas possíveis a um dado quesito num questionário de uma pesquisa de opinião foram mensuradas segundo uma escala de tipo *Likert*, resultando na seguinte tabela

Resposta	Freqüência	Proporção	Freq. Acumulada	Prop. Acumulada
DISCORDO FORTEMENTE	60	0,30	60	0,30
DISCORDO	80	0,40	140	0,70
NÃO DECIDIDO	30	0,15	170	0,85
CONCORDO	20	0,10	190	0,95
CONCORDO FORTEMENTE	10	0,05	200	1,00
Total	200	1,00	200	1,00

Note-se que o processo de acumulação pode ser realizado em qualquer sentido, indiferentemente.

Quando a acumulação é feita pela distribuição de percentuais, à percentagem acumulada até uma dada ordem dá-se o nome de percentil. Assim, usando o exemplo anterior, podemos dizer que as pessoas respondendo "não decidido" aquele item do questionário se localizam no 85º percentil da distribuição das respostas. Formalmente o percentil correspondente à $i^{ésima}$ ordem, pode ser notado como

$$Pi = 100 Fi N$$

onde Pi é o percentil e Fi é a frequencia acumulada até a iésima classe e N é o tamanho da amostra (total de observações).

Generalizando, a tabela de freqüências acumuladas tema a forma

Ordem	**Freqüência**	**Freqüência Acumulada**
01	f1	F1=f1
02	f2	F2=f1+f2
0m	fm	Fm=i=1mfi=N
TOTAL	N	N

I.3.3- Medidas Adequadas ao Nível Intervalo e Razão

Exatamente como nos casos de variáveis mensuradas aos níveis nominal e ordinal, é também adequado ao nível intervalo a sumarização dos dados através de uma tabela de freqüências. Para isso torna-se necessário naturalmente o estabelecimento de "classes operacionais", uma vez que o nível intervalo de mensuração não é caracterizado pela existência de "classes" "naturais" (caso das "categorias" e das "ordens" dos níveis de mensuração inferiores).

Assim, o primeiro passo para a tabulação será o cálculo de "intervalos de classe", nos quais ordenaremos os valores da variável a ser tabulada. Para isso estabelece-se primeiramente o número de classes em que particionaremos a variável. Esse processo de estabelecimento do número de classes é subjetivo e arbitrário; no entanto, deve-se evitar não só a partição num número demasiado de classes (o que seria econômico quando se possui um número elevado se observações) como também a partição em um número insuficiente de classes, o que deformaria a distribuição. É usual se usar para o cálculo do número de classes a seguinte relação empírica

$$n=1+3{,}3\log 10N$$

denominada fórmula de *Sturges*, onde *N* é o numero de unidades a analisar e *n* o numero de classes a estabelecer. Decorrente dessa formula temos a tabela de *Sturges* para o número de intervalos de classe *n* em um conjunto de dados (população ou amostra) de *N* observações

N	***n***
6 a 11	4
12 a 22	5
23 a 45	6
46 a 90	7
91 a 181	8
182 a 362	9
363 a 724	10
725 a 1448	11
1449 a 2896	12

O cálculo seguinte para a determinação dos intervalos de classe é o do tamanho dos intervalos. Para isso calcula-se inicialmente a amplitude da distribuição, entendida como sendo a diferença entre o maior e o menor valor da variável. Assim, o tamanho i do intervalo de classe pode ser obtido $i = \frac{\text{amplitude}}{n}$
onde *n* é o numero de intervalos de classe. Se o i calculado tiver casas decimais, é aconselhável o arredondamento para o primeiro número inteiro consecutivo superior.

Define-se ponto médio de classe como o número que cai exatamente na metade do intervalo de classe. É definido como

$$\text{ponto médio de classe} = \text{limite superior de classe} - 0{,}5i = \text{limte inferior de classe} + 0{,}5i$$

Segue-se implicitamente das observações acima que normalmente usa-se classes com intervalos iguais, embora tecnicamente isso não seja necessário e mesmo, em certas situações, desejável.

Tendo-se determinado os intervalos de classe, resta-se estabelecer os limites de cada classe. Nesse ponto é comum alguma confusão uma vez que o que se define como limite real de classe é por vezes diferente do limite de classe aparente. O limite real de classe determina os valores de inclusão nos intervalos de classe; é normalmente definido como sendo

Limite real inferior = limite aparente inferior - 0,5

Limite real superior = limite aparente superior + 0,5

e destinam-se para os casos de variáveis de tipo continuo, ou seja, que possuem casos decimais.

Exemplificando, suponhamos a seguinte tabela de freqüências:

Salário (cruzeiros) (limites aparentes)	**Limites Reais**	**N° de Empregados**	**Percentagens**	**Percentis**
50 a 59	49,5 a 59,5	8	12	12
60 a 69	59,5 a 69,5	10	16	28
70 a 79	69,5 a 79,5	16	24	52
80 a 89	79,5 a 89,5	14	21	73
90 a 99	89,5 a 99,5	10	16	89
100 a 109	99,5 a 109,5	5	8	97
110 a 119	109,5 a 119,5	2	3	100
Total		65	100%	100%

onde n=7 $i = \frac{119-50}{7} = \frac{69}{7} \cong 10$

As representações gráficas utilizadas ao nível nominal e ordinal podem também ser usadas com variáveis mensuradas ao nível de intervalo ou razão.

I.3.4- Medidas de Forma de Distribuição

Como já foi dito anteriormente, o objetivo da Estatística Descritiva é obter medidas que sumarizem ou descrevam sucintamente as distribuições que queremos estudar. Ao nível intervalo ou razão de mensuração podemos além de descrever as distribuições em termos de razões e proporções, podemos ainda obter medidas que descrevam sumariamente as formas que tomam essas distribuições. Um primeiro tipo dessas medidas é constituído pelas chamadas medidas de tendência central, consistindo em estatísticas que tentam capturar valores "típicos" representativos de toda a distribuição. Vejamos sistematicamente as medidas de tendência central mais usuais.

I.3.4.1- Medidas de Tendência Central

As medidas de tendência central mais utilizadas são a moda, a mediana e a média.

a. A Moda: é o correspondente ao nível intervalo da "categoria modal", vista anteriormente para o caso nominal. É definida simplesmente como o valor mais freqüente. Para o cálculo da moda a partir de dados agrupados utiliza-se a fórmula

$$Mo=LMo+f+1f-1+f+1i$$

onde Mo é a Moda,

LMo é o limite inferior da classe modal,

f+1 e f-1 são respectivamente as frequencias das classes acima e abaixo da classe modal e

i é o tamanho do intervalo de classe.

Assim, para o exemplo anterior

$$Mo=69,5+1410+1410=75,3$$

Observe-se que o cálculo da Moda é influenciado pelo agrupamento arbitrário feito, estando, portanto, sujeita a flutuações dependentes da maneira pela qual os dados foram agrupados. Além disso, pode ocorrer o caso em que duas (ou mais classes) tem idêntica freqüências, sendo essas as maiores na distribuição. Nesse caso diz-se que a distribuição é "bimodal".

(b) A Mediana – a mediana também corresponde ao nível de intervalo a em conceito já visto ao nível ordinal de mensuração. É definida como o 50ésimo percentil, ou seja, aquele ponto que divide a distribuição em duas metades de igual freqüência. Existe também uma fórmula para o cálculo de mediana a partir de dados agrupados:

$$Md=Lmd+N2-F-1fmdi$$

onde Md é a mediana, Lmd é o limite inferior do intervalo que contém a mediana,

N é o tamanho da amostra (total de observações)

F-1 é a freqüência acumulada até a classe anterior à que cai a mediana,

fmd é a freqüência da classe que contém a mediana e

i é o tamanho do intervalo de classe.

Assim, para o exemplo anterior:

$$Md=69,5+652-181610=78,5$$

(c) A Média - esta é a mais comum das medidas de tendência central. É definida como sendo a soma dos valores de distribuição dividida pelo número total de casos envolvidos. Formalmente

$$X=X1+X2++XNN=i=1NXiN$$

onde *i* representa o iésimo valor na distribuição.

A média tem duas propriedades algébricas importantes. A primeira delas é que a soma dos desvios em relação à média é zero. Algebricamente

$$i=1NXi-X=0$$

A prova dessa propriedade é bastante simples:

$$i=1NXi-X=i=1NXi-i=1NX$$

como X é uma constante

$$i=1NX=NX=Ni=1NXiN=i=1NXi$$

logo

$$i=1NXi-X=i=1NXi-i=1NXi=0$$

A segunda propriedade pode ser verbalizada da seguinte maneira: a soma dos desvios quadráticos em relação à média é menor que a soma dos desvios quadráticos em relação a qualquer outro número. Esta é a propriedade de "mínimos quadrados" da media, que pode ser escrita como:

$$i=1NXi-X2=mínimo$$

A prova dessa propriedade da média requer conhecimento de Cálculo e não será feita aqui. A quantidade definida acima, $i=1NXi-X2$ nos será de importância adiante, quando a usamos como uma medida de variação total.

A fórmula para o cálculo da média a partir de dados agrupados pode ser escrita da seguinte forma

$$X=i=1xifiN$$

onde xi é o ponto central da $i^{ésima}$ classe,

fi e a freqüência de $i^{ésima}$ classe,

e N é o tamanho da amostra (total de observações).

Observe-se que podemos escrever a fórmula acima em termos de freqüências relativas (proporções) fazendo

$$pi=fiN$$

Assim,

Utilizando o exemplo anteriormente visto, temos

Salários (Limites)	x_i	f_i	$p_i=f_i/N$	x_ip_i
54,5 a 59,5	54,5	8	0,12	6,54
59,5 a 69,5	64,5	10	0,16	9,72
69,5 a 79,5	74,5	16	0,24	17,88
79,5 a 89,5	84,5	14	0,21	17,75
89,5 a 99,5	94,5	10	0,16	15,12
99,5 a 109,5	104,5	5	0,08	8,38
109,5 a 119,5	114,5	2	0,03	3,44
Total	-	65	1	78,83

$$\therefore X=78,8$$

Nota :

1. Observe-se que a fórmula acima, desenvolvida para o caso em que os dados estão previamente agrupados (situação em que o valor calculado para a média é um valor aproximado, uma vez que estamos fazendo todos os valores dentro de cada classe serem iguais a x_i), é perfeitamente geral, uma vez que sempre podemos definir as classes (i.e. x_i) de tal forma a corresponderem a todos os valores observados. Ou seja, cada x_i pode representar um valor diferente observado em nosso conjunto de dados. Por exemplo, suponhamos os seguintes valores:

$$X: 2,1,2,2,3$$

Podemos calcular a média da forma

x_i	f_i	$p_i=f_i/N$	x_ip_i
1	1	0,20	0,20
2	3	0,60	1,20
4	1	0,20	0,60
Total	5	1	2,00

$$\therefore X=i=1Nxipi=2$$

Essa forma de expressar a média é a mais usual em Estatística. Retornaremos a ela com freqüência daqui em diante.

2. É usual a distinção, a nível de notação, entre a média da População e a da Amostra. A primeira é normalmente notada X e a segunda geralmente apresenta uma barra sobre o nome da variável (i.e. X).

(d) Comparação da Mediana com a Média –

A média e a mediana são duas estatísticas que descrevem satisfatoriamente o que poderíamos tomar como o valor "típico" de uma distribuição. A questão que surge é naturalmente, quando usar uma ou outra medida?

Uma importante característica da media, sua relativa estabilidade (comparada a uma mediana) ao longo das diferentes amostras, a tornam uma medida preferível à mediana. Existe uma situação, no entanto, em que a mediana nos fornece uma medida claramente mais "típica" da distribuição do que a média: é o caso em que a distribuição é fortemente assimétrica, ou seja, tem uma "cauda" alongada seja pra direita, seja pra esquerda. A mediana, por não depender de valores extremos da distribuição, não varia com o nível de assimetria da mesma, enquanto que a média é altamente sensível a valores extremos. Por isso, quando a distribuição for fortemente assimétrica é preferível se utilizar a mediana ao invés da média. Essa é a razão pela qual quando caracterizamos distribuições de rendimentos é mais usual a referência a mediana da distribuição do que a média da distribuição dos rendimentos.

De fato, a relação entre a média e a variância nos indica claramente a direção e a intensidade da assimetria da distribuição. Quando a mediana tem valor superior ao da média, a distribuição mostra uma assimetria negativa, ou seja, uma cauda alongada para a esquerda. No caso inverso a assimetria é dita positiva. Quanto maior a diferença entre média e mediana, maior o nível de assimetria na distribuição.

Quando a distribuição é simétrica, média e mediana apresentam valores idênticos.

I.3.4.2- Medidas de Dispersão

Para descrever uma distribuição necessitamos também o nível de dispersão (ou concentração) que a caracteriza. Em outras palavras, precisamos conhecer a heterogeneidade ou variação nos dados com que estamos trabalhando. Entre outras coisas, essa medida de dispersão nos ajudará a estabelecer o nível de "representatividade" ou "tipicidade" das medidas de tendência central. Um exemplo ajudará a esclarecer o ponto; suponhamos as seguintes distribuições:

Distribuição X	
X	$f_i(X)$
0	0
1	0
2	0
3	2
4	6
5	22
6	10
7	0
8	0
9	0
10	0
Total	40

Distribuição Y	
Y	$f_i(Y)$
0	3
1	15
2	0
3	0
4	0
5	7
6	0
7	0
8	0
9	0
10	15
Total	40

$\bar{X}=5$ $\bar{Y}=5$

Observando as duas distribuições acima vemos claramente que na distribuição *X* a média é um valor representativo da distribuição enquanto que na distribuição *Y* a média tem escasso valor substantivo. Isto porque a "dispersão" das observações em torno da média na primeira distribuição é "pequena" enquanto que na segunda é "grande" essa dispersão. Assim, a interpretação substantiva das medidas de tendência central de uma dada distribuição depende da magnitude da "dispersão" dessa mesma distribuição.

Como no caso das medidas de tendência central, existem várias medidas de dispersão. Vejamos as mais usuais e importantes.

a. A Amplitude – a amplitude é a mais simples das medidas de dispersão. Como já foi visto, é definida como a diferença entre o maior e o menor valor na distribuição.

No caso das distribuições acima, as amplitudes são respectivamente:

Amp.de $X=6-3=3$

Amp.de $Y=10-0=10$

A amplitude é, por razões obvias (ela só nos diz onde a distribuição começa e onde acaba), uma medida muito rudimentar. As medidas de dispersão mais comumente usadas são aquelas que usam as medidas de tendência central em suas definições. Examinaremos essas medidas a seguir

b. Variância e Desvio Padrão

A medida de tendência central mais utilizada é como já dissemos antes, a média. Assim, podemos naturalmente tomá-la como ponto de referência e analisar o nível de dispersão dos valores na distribuição em relação a ela. Ou seja, podemos tomar como base o desvio em relação à média

$$Xi-X$$

representando a diferença (ou distância) entre observação *i* e a média da distribuição X, como base para se definir as medidas de dispersão.
Para caracterizar o nível de dispersão em toda a distribuição devemos, naturalmente, agregar todos os desvios, ou seja, fazer

$$i=1NXi-X$$

Mas, como já vimos antes, uma das propriedades da média é justamente fazer com que a expressão acima seja igual a zero, o que a desqualificaria como medida de dispersão. Para solucionar esse problema, poderíamos tomar os desvios não em termos de valores relativos (i. e. com sinal), mas em termos de valores absolutos

$$i=1NXi-X$$

Instintivamente, a medida acima captura adequadamente o nível de dispersão nos dados. Entretanto, a operação com valores absolutos apresenta sérias dificuldades para o tratamento matemático (por exemplo, a expressão acima não é diferenciável), o que a invalida como uma medida desejável. Uma medida de dispersão que não apresenta esse tipo de problema é

$$i=1NXi-X2$$

Essa é, portanto uma medida adequada da dispersão ou variação total na distribuição, representando a soma dos desvios quadráticos em relação à média. Como vimos antes, uma das propriedades da média é exatamente fazer com que essa variação total seja a menor possível (a chamada propriedade de "mínimos quadrados" da média).

A variação total da distribuição, no entanto, é uma função direta do número de observações que compõem a distribuição: quanto maior o N, maior a variação total. Isto é evidentemente uma propriedade indesejável para uma medida de dispersão, uma vez que, por exemplo, invalidaria a comparação das variações entre duas ou mais distribuições. O que interessa, naturalmente, é uma variação ou desvio médio, um que pudesse ser qualificado de "desvio típico" da distribuição. Assim podemos fazer

$$SX2=i=1NXi-X2N$$

A essa medida, interpretável como o desvio quadrático médio, damos o nome de Variância. Como nessa medida os desvios estão elevados ao quadrado, podemos extrair

a sua raiz quadrada para obtermos os desvios na mesma escala dos valores da distribuição. Ou seja, podemos definir

$$SX=i=1NXi-X2N$$

medida que é chamada de Desvio Padrão.

Quando trabalhamos com dados agrupados, a fórmula da variância é

$$SX2=i=1Nxi-X2fiN=i=1Nxi-X2pi$$

onde x_i, f_i e p_i tem a mesma definição que anteriormente. Naturalmente para se obter o Desvio Padrão extraímos a raiz quadrada da expressão acima

$$SX=i=1Nxi-X2pi$$

Utilizando os exemplos anteriores, podemos calcular:

Distribuição X					
X	$f_i(X)$	pi=fiN	xi-X	xi-X2	xi-X2pi
3	2	0,05	-2	4	0,20
4	6	0,15	-1	1	0,15
5	22	0,55	0	0	0,00
6	10	0,25	1	1	0,25
Total	40	1,00	-	-	0,60

$$X=5 \qquad SX2=0,60 \qquad SX=0,78$$

Distribuição Y					
Y	$f_i(Y)$	pi=fiN	yi-Y	yi-Y2	yi-Y2pi
0	3	0,08	-5	25	2,00
1	15	0,38	-4	16	6,08
5	7	0,16	0	0	0,00
10	15	0,38	5	25	9,50
Total	40	1,00	-	-	17,58

$$Y=5 \qquad SY2=17,58 \qquad SX=4,19$$

Notas: (1) Tal como é feito em relação à média, costuma-se distinguir a nível notacional, a Variância (e o desvio-padrão) da população e da amostra. Para a população utiliza-se a letra grega para designar o Desvio-padrão e 2 para designar a variância. Para a amostra os símbolos usados são S2 e S para a variância e desvio-padrão, respectivamente.

(2) Devido à questões relacionadas com a inferência estatística (o que será detalhado posteriormente), a variância da Amostra tem como denominador *N-1*, e não mais *N* como é o caso da Variância da População. Assim a variância (e conseqüentemente o desvio-padrão) é definida como

$$SX2=i=1NXi-X2N-1$$

Obviamente, a diferença de fórmula entre o parâmetro da população e a sua estimativa na amostra diminui conforme N cresce, sendo absolutamente irrelevante para amostras grandes (ou seja, para N's grandes).

Um quadro sumário destas diferenças de notação poderia ser:

Distribuição	Estatística		
	Média	**Variância**	**Desvio-Padrão**
População	X=XiN	X2=i=1NXi-X2N	X=X2
Amostra	X=XiN	SX2=i=1NXi-X2N-1	SX=SX2

I.3.5 – Transformações Lineares

Transformações lineares são mudanças (recodificações) que se aplicam em variáveis e que tem a seguinte forma geral

Xi'=a+bXi (onde a e b são constantes)

que corresponde a equação da linha reta. Examinemos esse tipo de transformação em mais detalhe.

a. Mudança de Origem – Transformações do tipo

$$Xi'=a+Xi$$

(ou seja, b=1), correspondem a uma mudança na origem da escala em que X esta mensurado. Esse tipo de transformação tem as seguintes conseqüências em termos das estatísticas que descrevem a distribuição:

(a.1) Média: X'=a+XN

$$X'=a+XN=1Na+X=1Na+X$$

$$=aN+XN=NaN+X$$

$$=a+X$$

Ou seja, a média de *X'* é igual a média de *X* acrescida da constante a.

(a.2) Variância: SX'2=X'-X'2N-1

Como X'=a+X e X'=a+X

$$SX'2=a+X-a+X2N-1=a+X-a-X2N-1=X-X2N-1$$

$$\therefore SX'2=SX2$$

Ou seja, a dispersão dos dados em relação à média não se altera.

b. Mudança de Escala – transformações do tipo

$$Xi'=bXi$$

(ou seja, a=0), correspondem a uma mudança na unidade de mensuração que compõe a escala, significando seja uma compressão da escala (quando b<1), seja um espichamento dela (quando b>1). Suas conseqüências em termos das estatísticas que descrevem a distribuição são as seguintes

(b.1) Média: X'=bXN

X'=bXN=bXN=bX

Ou seja, a média de *X'* é igual a b vezes a média de X.

(b.2) Variância: SX'2=bX-bX2N-1

SX'2=bX-bX2N-1=bX-X2N-1

=b2X-X2N-1=b2X-X2N-1

∴SX'2=b2SX2

Ou seja, a variância de *X'* é b2 vezes a variância de X.

c. Transformação Linear Geral – quando a0 e b1, temos as seguintes conseqüências
(cujas deduções são paralelas as feitas acima):

(c.1)

(c.2)

d. Padronização – uma transformação linear importante em Estatística é a chamada padronização. A transformação efetuada na Padronização é a seguinte (para o caso de Amostras):

=1SXXi-XSX=bXi+a

onde então b=1SX e a=-XSX.
Usando as propriedades da transformação linear geral temos que

(d.1) Z=a+bX=-XSX+1SXX=-XSX+XSX=0

∴Z=0

(d.2) SZ2=b2SX2=1SX2SX2=1SX2SX2

$$\therefore SZ2 = 1$$

Assim, variáveis que foram padronizadas tem como propriedades

Esse tipo de transformação será de grande importância quando estudarmos os testes de hipóteses.

CAPÍTULO 2 - ASSOCIAÇÃO: RPE E DESVIO DA INDEPENDÊNCIA

2. Descrevendo relações entre 2 (ou mais) variáveis. Diferenças proporcionais. Conceitos de Associação: independência e Redução Proporcional do Erro de predição (RPE).

2.1 Tabelas de Contingência

Vamos examinar agora a questão de como podemos descrever a relação entre duas ou mais variáveis. No presente capítulo estaremos limitados às relações entre variáveis qualitativas, isto é, variáveis medidas no nível Categórico ou no nível Ordinal, deixando para os capítulos seguintes as questões envolvendo variáveis numéricas, isto é, variáveis medidas no nível de Intervalo ou de Razão.

Devemos lembrar que, como vimos no capitulo anterior, no nível qualitativo de mensuração, a operação básica é a de **contagem das freqüências** de ocorrência das diversas categorias entre nossas observações. Assim, utilizando os dados hipotéticos de nossa turma de aula (distribuídos em classe), podemos, por exemplo, obter as distribuições de freqüência da variável sexo e das respostas à questão sobre se "existe discriminação contra mulheres no Brasil". Os resultados dessa operação seriam:

Tabela 2.1: Sexo

	Frequency	Percent	Valid Percent	Cumulative Percent

Valid	Masculino	9	45,0	45,0	45,0
	Feminino	11	55,0	55,0	100,0
	Total	20	100,0	100,0	

Tabela 2.2: Existe Discriminação contra mulheres no Brasil?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Sim	12	60,0	63,2	63,2
	Não	7	35,0	36,8	100,0
	Total	19	95,0	100,0	
Missing	Sem opinião/Não sei	1	5,0		
Total		20	100,0		

Como podemos ver, dos 20 alunos em nossa classe hipotética, 9 são do sexo masculino e 11 do sexo feminino. Descrevendo em termos percentuais, constatamos que a freqüência relativa de mulheres é de 55%, enquanto que os homens constituem os 45% restantes.

Por outro lado, quando olhamos para as resposta à questão da discriminação contra mulheres, observamos que 1 dos alunos, por alguma razão, não respondeu à questão, sendo então considerado "Falta de Informação" ("Missing Data"). Com isso, nosso total de casos fica reduzido a 19 observações. Dessas, 12 concordam com a afirmativa de que sim, existe discriminação contra mulheres, ao passo que 7 discordaram da afirmativa. Em termos relativos, entre os casos válidos, 63,2% concordam com a afirmativa, com os restantes 36,8% discordando dela.

Quando temos em nosso conjunto de dados, como é o caso de nossa turma hipotética, mais de uma variável coletada, podemos analisar a distribuição simultânea de duas (ou mais) variáveis conjuntamente. Por exemplo, podemos obter a distribuição conjunta das duas variáveis acima, contando a ocorrência de todas as combinações possíveis das categorias das duas variáveis. A representação tabular dessa contagem de freqüências de ocorrência das categorias combinadas denomina-se "tabela de contingência" ou "tabela cruzada", representando essa tabulação a relação entre as variáveis examinadas. No exemplo das variáveis selecionadas acima, a tabela de contingência relacionando sexo e opinião sobre discriminação contra mulheres teria a seguinte forma:

Tabela 2.3 :Existe Discriminação contra mulheres no Brasil? * Sexo Crosstabulation

			Sexo		
			Masculino	Feminino	Total
Existe Discriminação contra mulheres no Brasil?	Sim	Count	2	10	12
		% of Total	10,5%	52,6%	63,2%
	Não	Count	6	1	7
		% of Total	31,6%	5,3%	36,8%
Total		Count	8	11	19
		% of Total	42,1%	57,9%	100,0%

Observe-se que, nesse caso, como temos 1 caso de falta de informação na questão sobre discriminação, o total conjunto de respostas válidas se reduz a 19, alterando assim a distribuição relativa da variável sexo. Como o respondente que não deu resposta à questão da discriminação era homem, a distribuição da variável sexo se altera para 8 homens (ou seja, 42,1% do total) e 11 mulheres (57,9%). Podemos ver também que, dos 19 respondentes válidos, 10 são mulheres que concordam com a afirmativa (ou seja, 52,6% do total), ao passo que, por exemplo, 6 (ou 31,6%) são homens que discordam dessa afirmativa.

Como já foi indicado, podemos construir tabelas de contingência, descrevendo inter-relações, envolvendo um número maior de variáveis simultaneamente. Por exemplo, novamente com os dados hipotéticos de nossa turma de aula, podemos adicionar à tabela de contingência acima mais uma (ou ainda mais) variável. Se nossa escolha fosse a variável Cor do aluno, a tabela de cruzada tri-dimensional resultante teria a forma:

Tabela 2.4: Sexo * Existe Discriminação contra mulheres no Brasil? * Cor Crosstabulation

			Existe Discriminação contra mulheres no Brasil?		
Cor			Sim	Não	Total
Branca	Sexo	Masculino	1	3	4
		Feminino	5	0	5
Preta	Sexo	Masculino	0	0	0
		Feminino	1	1	2
Parda	Sexo	Masculino	1	3	4
		Feminino	4	0	4

Como podemos ver, a tabela acima reporta todas as combinações simultâneas entre as categorias das 3 variáveis tomadas conjuntamente. Com isso, observamos, por exemplo, que 4 mulheres pardas declararam concordar com a afirmativa sobre a existência de discriminação, enquanto que foram 3 os homens brancos que discordaram dessa afirmativa.

Na tabela acima, temos relacionadas três características distintas simultaneamente e, por isso, a tabela de contingência é dita tri-dimensional. Evidentemente, podemos construir uma tabela com o número de dimensões (variáveis) que desejarmos. No entanto, a análise dos dados cresce em complexidade conforme aumentamos o número de dimensões, não sendo freqüentes, portanto, análises envolvendo um número grande de variáveis, digamos mais de seis ou sete características simultaneamente. O caso mais simples, as chamadas tabelas bi-dimensionais, será objeto das seções que se seguem, nos permitindo desenvolver a discussão sobre as diferentes noções/definições de "associação" entre variáveis. Os casos mais complexos, envolvendo mais de duas dimensões são objeto de um curso específico, Lego III. Antes de prosseguirmos, porém, vejamos algumas questões de nomenclatura, preparando uma linguagem comum que nos permita desenvolver a nossa discussão sobre a descrição da associação entre variáveis.

2.2 Nomenclatura

Como já foi indicado, estamos interessados em estudar relações entre duas ou mais características medidas ao nível qualitativo (nominal ou ordinal). A forma usual de se representar uma relação entre variáveis categóricas é através da contagem de freqüências de co-ocorrências de todas as combinações possíveis de categorias, escritas sob a forma de uma tabela de contingência. No caso mais simples em que temos apenas duas variáveis, essa tabela toma a seguinte forma de distribuição de freqüências conjuntas ou "tabela de contingência":

A: variável Linha (i)	**B: Variável Coluna (j)**				**Total da Variável Linha**
	1	**2**	**.....**	***c***	
1	n_{11}	n_{12}	…	n_{1c}	n_{1+}
2	n_{21}	n_{22}	…	n_{2c}	n_{2+}
.	.	.	…	.	.
.	.	.	…	.	.
.	.	.	…	.	.
l	n_{l1}	n_{l2}	…	n_{lc}	n_{l+}
Total da Var Coluna	n_{+1}	n_{+2}	…	n_{+c}	n_{++}

onde n_{ij} indica a freqüência conjunta da categoria i da variável linha A e da categoria j da variável coluna B.

As distribuições de freqüência de cada uma das variáveis envolvidas, denominadas freqüências marginais (por se encontrarem em cada uma das duas margens da tabela), são notadas como

$$ni+ = j=1cnij \quad \text{e} \quad n+j= i=1lnij$$

O total geral sendo

$$n++ = i=1lj=1cnij = i=1lni+ = j=1cn+j = N$$

Assim, por exemplo, na tabela 2.3, temos que $n_{11} = 2$, $n_{12} = 6$, $n_{21} = 10$ e $n_{22} = 1$, do que resulta que as distribuições marginais (i. e. univariadas) são

$$n_{1+} = j=12n1j = n_{11} + n_{12} = (2+6) = 8.$$

De forma semelhante, por exemplo, temos que,

$$n_{+2} = i=12ni2 = n_{12} + n_{22} = (6+1) = 7$$

e

$$n_{++} = n_{11} + n_{12} + n_{21} + n_{22} = (2+6+10+1) = 19.$$

2.3: O Conceito de Independência e a estatística X^2

O objetivo de se construir uma tabela de contingência é, usualmente, o de examinar a eventual relação entre as variáveis em que estamos interessados. Por exemplo, na tabela 2.3 podemos estar interessados em estudar a relação entre o sexo do aluno e a sua opinião sobre a existência de discriminação contra as mulheres no Brasil, possivelmente a partir de um modelo teórico que postule um efeito da primeira variável sobre a segunda. Podemos supor, por exemplo, que as mulheres sejam mais propensas a perceber uma discriminação contra elas do que os homens o seriam.

Uma maneira tradicional de se analisar as relações em uma tabela de contingência envolve o calculo das distribuições relativas (e.g. percentuais) ao longo de cada categoria da variável que tomamos como explicativa da outra variável. No nosso exemplo, tomamos o sexo do aluno como explicação para eventuais diferenças na opinião sobre discriminação contra mulheres. Assim, calcularemos a distribuição das freqüências relativas da variável opinião dentro de cada categoria de sexo (formando então a chamada **distribuição cond**icional de opinião **dado** o sexo do respondente) e compararemos os resultados correspondentes entre as categorias de sexo. Expressando essa operação de cálculo das distribuições condicionais para a tabela 2.3 em termos percentuais, obtemos os seguintes resultados:

Tabela 2.5 Distribuições Condicionais de Opinião sobre Discriminação contra mulheres dado o sexo do respondente

		Sexo		
		Masculino	Feminino	Total
Existe Discriminação contra mulheres no Brasil?	Sim	25,0%	90,9%	63,2%
	Não	75,0%	9,1%	36,8%
Total		100,0%	100,0%	100,0%

Observe que, para o conjunto dos alunos, 63,2% concordam com a afirmativa sobre discriminação e 36,8% discordam dessa afirmação. No entanto, entre os homens apenas 25,0% deles concordam, ao passo que a grande maioria, numa extensão de 75,0% discordam da afirmativa. Simetricamente, entre as mulheres, nada menos de 90,1% delas concordam enquanto que apenas 9,1% discordam. Com isso, podemos ver que, nesses dados, as mulheres mostram, de fato, uma maior propensão a apoiar a afirmativa de que existe discriminação do que os homens.

Observe-se também que, no caso dessa nossa tabela 2.5 temos apenas 1 diferença percentual entre homens e mulheres, dado que cada uma das duas distribuições

condicionais tem soma 100%, fazendo com que a soma das diferenças seja sempre igual a zero. Com isso, as diferenças entre os que dizem "sim" e os que dizem "não" são idênticas, com sinais trocados: entre homens e mulheres a diferença na propensão a discordar é de Δ = (75,0% - 9,1%) = 65,9%; enquanto que a mesma diferença na propensão a concordar é de Δ = (25,0% - 90,9%0 = -65,9%. Em outras palavras, numa tabela em que ambas as variáveis são dicotômicas, isto é, têm apenas 2 categorias cada, temos apenas 1 diferença percentual entre as distribuições condicionais.

O conceito de Independência entre duas variáveis parte do princípio de que, em sua ocorrência, a variável explicativa não afetará a distribuição da variável explicada. No nosso exemplo, vimos que 63,2% dos alunos concordam com a afirmação sobre discriminação. Assim, se o sexo do aluno não afetasse a sua opinião, deveríamos observar esse mesmo percentual de pessoas concordando com a afirmativa tanto entre homens como entre mulheres. Nesse caso, diríamos que opinião e sexo são Independentes. Caso, contrário, quando essa condição não se verifica, então diríamos que as duas variáveis são Associadas. Ou seja, Associação aqui é entendida como "desvio do caso de Independência". Nesse sentido, os dados da tabela 2.5 indicam uma situação de Associação.

Em termos formais, podemos escrever a seguinte definição de Independência: duas variáveis categóricas serão independentes se, para todos os valores de i e de j

$$n_{ij} / n_{+j} = n_{i+} / n_{++}$$

do que decorre , definindo alternativamente,

$$n_{ij} = F_{ij} = n_{i+}\, n_{+j} / n_{++} \quad \text{(eq. 2.1)}$$

Observe que a equação acima nos fornece um contrafactual: expressa a freqüência que devemos esperar se as duas variáveis fossem independentes, especificando o chamado "modelo de Independência". Por exemplo, se as duas variáveis na tabela 2.3 fossem independentes , então deveríamos esperar que

$$n_{11} = F_{11} = n_{1+}\ n_{+1} / n_{++}$$

o que, numericamente, significa que esperaríamos a seguinte freqüência nessa célula:

$$F_{11} = (8 \text{ X } 12) / 19 = 5,1$$

Evidentemente podemos efetuar calculo similar para as demais células da tabela, obtendo o seguinte quadro de Freqüências Esperadas no caso de independência entre as variáveis:

Tabela 2.6 : Frequencias Esperadas no caso de independência entre Sexo e Opinião sobre discriminação contra mulheres no Brasil

	Sexo	Total

		Masculino	Feminino	
Existe Discriminação contra mulheres no Brasil?	Sim	5,1	6,9	12,0
	Não	2,9	4,1	7,0
Total		8,0	11,0	19,0

Entendendo Associação como "desvio da independência", podemos cotejar essa tabela que expressa as freqüências esperadas no caso das duas variáveis serem independentes (tab. 2.6) com a tabela obtida por observação dos dados (tab. 2.3). Quanto mais elas discreparem, maior será, ceteris paribus, o nível de associação entre as duas varáveis. Pearson (1904) sugeriu uma estatística que sumariza essa discrepância, tendo a denominado "Qui-quadrado", sendo definida com a forma

$$X^2 = \text{i=1lj=1c nij –Fij 2/ Fij} \qquad \text{(eq. 2.2)}$$

onde n_{ij} representa as freqüências observadas (tab. 2.3) e F_{ij} as respectivas freqüências esperadas no caso de independência (eq 2.1). As diferenças são elevadas ao quadrado por razões semelhantes ao que acontece no caso do cálculo da Variância, isto é, a sua soma simples é sempre igual a zero.

Podemos então calcular X^2 para o nosso exemplo, fazendo

Célula	Freq. Obs. nij	Freq. Esper. Fij	nij –Fij 2	nij –Fij 2/ Fij
11	2	5,1	9,61	1,88
12	10	6,9	9,61	1,39
21	6	2,9	9,61	3,31
22	1	4,1	9,61	2,34
Soma	19	19	-	8,92

Um outro exemplo numérico tirado dos dados de nossa turma hipotética nos é dado pelo inter-relacionamento das variáveis Cor e Estrato Social. Essa é uma tabela bem mais complexa, envolvendo 3 categorias para cada variável, resultando numa tabela com 9 células. Nesse caso, pode-se demonstrar que existem (*l* -1) (*c*- 1) = (3-1) (3-1) = 4 diferenças percentuais possíveis nessa tabela. Esse cálculo será explicado e desenvolvido mais adiante, quando estivermos discutindo testes de significância. Fazendo esse cruzamento, obtemos

Tabela 2.7 : Estrato Social * Cor Crosstabulation

			Cor			
			Branca	Preta	Parda	Total
Estrato Social	A	Count	4	0	1	5
		% within Cor	44,4%	,0%	11,1%	25,0%
	B	Count	4	1	6	11
		% within Cor	44,4%	50,0%	66,7%	55,0%
	C	Count	1	1	2	4
		% within Cor	11,1%	50,0%	22,2%	20,0%
Total		Count	9	2	9	20
		% within Cor	100,0%	100,0%	100,0%	100,0%

Novamente, as distribuições condicionais discrepam das distribuições marginais, indicando a presença de Associação entre as variáveis. Por exemplo, enquanto que 11,1% dos Brancos encontram-se no denominado estrato C, as percentagens correspondentes entre Pardos é de 22,2% e entre Pretos é de 50,0%. Observações semelhantes podem ser feitas quanto às demais combinações de categorias. Com isso, podemos também calcular as freqüências esperadas no caso de independência (segundo a eq. 2.1). Nesse caso, obtemos em cada célula

Tabela 2.8: Frequencias Observadas e Esperadas no Caso de Independencia.

			Cor			
			Branca	Preta	Parda	Total
Estrato Social	A	Count	4	0	1	5
		Expected Count	2,3	,5	2,3	5,0
	B	Count	4	1	6	11
		Expected Count	5,0	1,1	5,0	11,0
	C	Count	1	1	2	4
		Expected Count	1,8	0,4	1,8	4,0
Total		Count	9	2	9	20
		Expected Count	9,0	2,0	9,0	20,0

Aplicando o calculo da estatística X^2, indicado pela equação 2.2 acima, obtemos o valor $X^2 = 4,25$.

2.4 Medidas de Associação

De um modo geral, concorda-se que uma medida de associação deve atingir o valor máximo de 1 para o caso de associação "completa" e 0 para o caso de independência (X^2). No caso das categorias envolvidas possuírem alguma ordenação entre si, então a medida de associação deve ser capaz de variar entre os limites de -1 para uma associação completa "negativa", e +1 para uma associação positiva, passando pelo ponto de mínimo 0, para o caso de independência entre as variáveis.

O problema é que uma série de fatores pouco tendo a ver com a associação entre as variáveis propriamente ditas, afeta a magnitude das medidas utilizadas. Assim, as medidas mais insensíveis a esses fatores de perturbação são normalmente preferíveis.

Como vimos anteriormente, a estatística X^2 mede o nível de desvio do caso empiricamente observado em relação ao padrão que observaríamos se as variáveis fossem independentes. Quanto maior a discrepância entre o observado e o esperado, maior o valor calculado de X^2. Assim sendo, poderíamos utilizar X^2 como um indicador de associação. Infelizmente, embora o limite inferior de X^2 seja o valor zero (caso de independência), o valor superior não está limitado ao teto da unidade, como seria desejável, uma vez que X^2 é função também do número de casos com que estamos trabalhando, crescendo conforme o N também cresce. Um exemplo simples mostra claramente esse fato:

Tabela A

3	2	5
2	3	5
5	5	10

$X^2 = 0,4$

Tabela B

30	20	50
20	30	50
50	50	100

$X^2 = 4,0$

O nível de associação nas duas tabelas acima é, obviamente o mesmo, apenas a tabela B tem dez vezes mais casos do que a tabela A. O valor de X^2 também segue essa proporção, sendo também dez vezes maior em B do que em A.

Para se obter, então, uma medida de associação a partir de X^2, devemos dividi-la pelo número de casos com que estamos trabalhando, o nosso N. Obtemos assim a medida denominada "coeficiente de contingência médio quadrático", grafada como ϕ^2 ("fi" quadrado):

$$\phi^2 = X^2 / N$$

Numa tabela 2X2 , ϕ^2 varia entre 0 e 1, sendo zero quando as variáveis são independentes e atingindo um no caso de associação perfeita. Nos exemplos vistos das tabelas A e B acima, os valores de ϕ^2 são idênticos, em ambos os casos tendo o valor aproximado de $\phi^2 = 0,04$.

No caso do nosso exemplo, envolvendo as variáveis "sexo" e "opinião sobre discriminação contra mulheres", obtivemos uma estatística $X^2 = 8,92$, a qual, com um N=19, implica num $\phi^2 = 8,92 / 19 = 0,47$.

Uma variante dessa medida é o chamado "Coeficiente de Contingência" quadrático, sugerido por Pearson (1904), dado por

$$P^2 = X2 /N1 + X2 /N = 21 + \Phi 2$$

Que, no caso do exemplo acima toma o valor de

$$P^2 = 0,47/ 1+ 0,47 = 0,32$$

$$P = P2 = 0,57$$

Os coeficientes Φ^2 e P, embora tenham a variação prevista no caso das tabelas 2X2, não podem atingir o valor 1 para o caso mais geral das tabelas *l X c*. Pode ser mostrado que, em geral, mesmo no caso de associação perfeita, o valor dessas estatísticas depende do número de linhas e de colunas da tabela, o que levou a novas modificações

nesses coeficientes para remediar essa deficiência, dando origem ao coeficiente T de Tschuprow e o coeficiente C de Cramér.

$$T = 2l\text{-}1(c\text{-}1)$$

Atingindo o limite inferior zero no caso de independência e o limite teórico de 1 no caso de tabelas quadradas. No caso em que *l ≠ c,* T terá limite superior menor que 1. Para efetuar mais uma correção Cramér propôs a medida

$$C = 2(l\text{-}1;c\text{-}1)$$

Onde o denominador é o menor valor entre *(l-1) e (c-1)* . C pode atingir o valor 1 para qualquer valor de *l* e de c no caso de associação perfeita entre as variáveis.

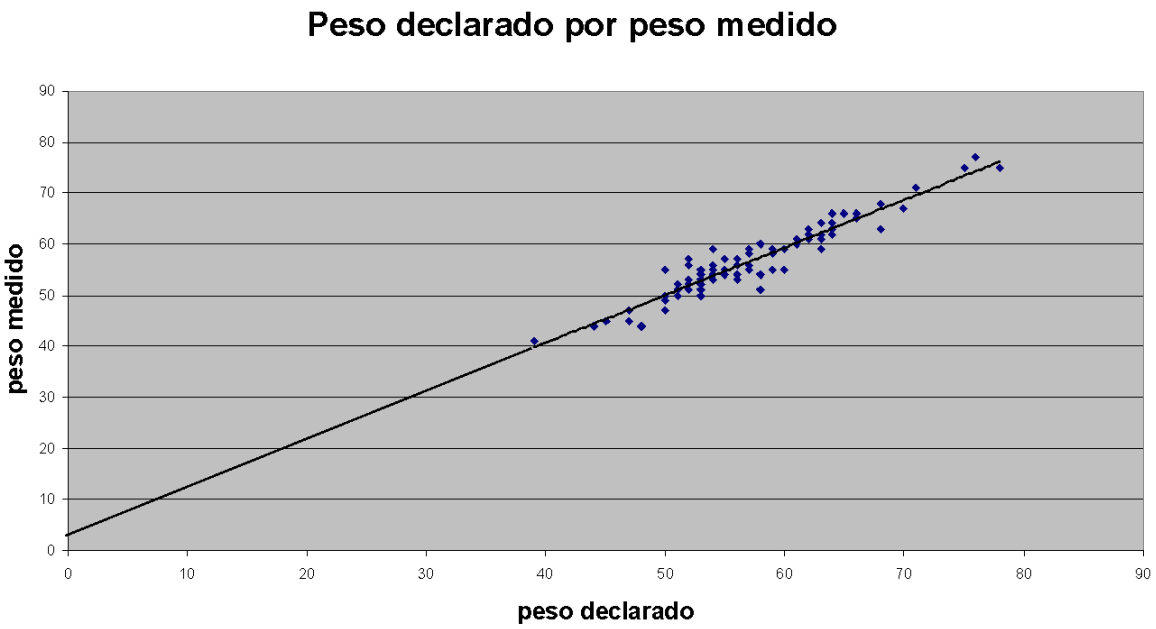
No nosso exemplo da tabela 2.3 , como l = c= 2, fazendo que os dois denominadores, de T e de C sejam iguais a 2. Ou seja,

$$T = C = 22 = 0{,}472 = 0{,}48$$

5. : Associação como Redução Proporcional no Erro de Predição

1. <u>Associação entre duas variáveis quantitativas</u>

<u>3.1) Regressão linear por mínimos quadrados</u>



$Y_i = A + BX$

incluindo o termo de erro:

$Y_i = A + BX + E_i$ (equação 1)
$= Y_i + E_i$

Esta idéia pode ser visualizada na figura abaixo:

A figura acima revela que:

$$E_i = Y_i - \hat{Y}_i = Y_i - (A + BX_i)$$

o resíduo (Ei) pode ser positivo ou negativo (acima ou abaixo da linha). Uma linha que se ajusta bem aos dados faz com que os resíduos sejam pequenos.

O somatório de todos os erros em torno de uma linha
ajustada aos dados será zero porque os valores negativos e positivos se anualarão entre si.

Na realidade qualquer linha

(equação 2)

produz o valor zero dos resíduos

$$\sum_{i=1}^{n} E_i = 0$$

$\bar{Y}_i = A + B\bar{X}_i$

Se substrairmos a equação 2 da 1 temos:

Então somando todas as observações temos que:

$$\sum_{i=1}^{n} (Y_i - \bar{Y}_i) - B\sum (X_i - \bar{X}) = 0 - b \times 0 = 0 \quad \text{(equação 3)}$$

Há duas alternativas:

(1) Encontrar A e B para minimizar os valores absolutos dos resíduos,

(2) ou, encontrar A e B para minimizar o quadrado dos resíduos.

Os quadrados são mais faceis de manipular matematicamente usaremos do que valores absolutos, e são utilizados mais frequentemente. Mas a regressão por "least absolute values" (LAV) também pode ser útil na medida em que é mais resistente a "outlying observations".

A "regressão por mínimos quadrados" minimiza a soma dos resíduos ao quadrado considerando todas as observações, ou seja, procuramos valores de A e B que mnimizem:

Os coeficientes da regressão por mínimos quadrados podem ser obtidos da seguinte forma:

$$A = \overline{Y} - B\,\overline{X}$$

$$B = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right.}{\sum\left(X_i - \overline{X}\right)^2}$$

Os dados sobre peso declarado e medido apresentados no primeiro "scatter plot" desta aula podem ser usados para exemplificar o calculo dos coeficientes da regressão:

n = 101

$$\sum \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right) = 4435$$

$$\overline{Y} = \frac{5780}{101} = 57{,}23$$

$$\sum \left(X_i - \overline{X} \right)^2 = 4539$$

$\overline{X} = 57{,}31$

A regressão por mínimos quadrados para estes dados sobre peso medido e declarado é:

$$peso\_medido = 1{,}79 + 0{,}9771 \times peso\_declarado$$

B = 0,977 significa que um aumento de 1 kg no peso declarado está associado em média à um aumento de 0,97 kg no peso medido.

A inteseção A é o valor ajustado de Y quando X = 0.

Erro padrão dos erros:

Além de calcular a "linha dos mínimos quadrados" é importante saber qual a proximidade da linha em relação aos pontos. Para tanto, podem ser utilizas a variância dos erros e "erro padrão" dos erros:

$$S_E^2 = \frac{\sum E_i^2}{n-2}$$

Correlação simples:

O coeficiente de correlação é uma medida relativa do ajuste:

Em que medida nossa predição de Y melhora quando baseamos esta predição na relação linear entre Y e X?

Uma medida relativa exige um ponto de comparação:

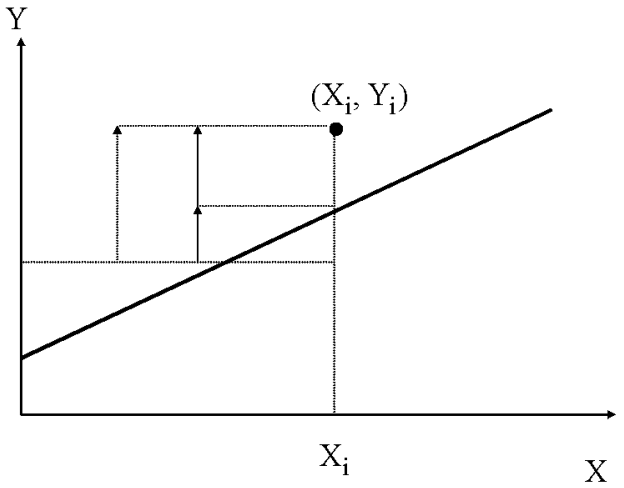
Em que medida Y pode ser previsto se X não for considerado?

Não levar em consideração X (variável independente) implica em estimar a equação:

$$Y_i = A' + E'$$

Os valores ajustados são constantes (não mudam). Para ajustar esta constante também podemos empregar o método dos mínimos quadrados para diminuir os erros.

O valor de A' que minimiza essa soma dos quadrados é simplesmente a média da variável dependente



3.2 Calculando o coeficiente de determinação (R^2)

Tendo em vista que a variação de Y deve-se tanto ao efeito de X quando ao erro aleatório, podemos partir a soma total dos quadrados em uma distribuição de escores entre um componente sistemático e outro aleatório.

Inicialmente crie um desvio subtraindo a média Y do valor observado Y_i. Em seguida adicione e diminua o valor predito pela regressão linear (Yi) deste desvio, produzindo assim a seguinte identidade:

$$Yi-Y=Yi-Y+Yi-Yi$$

Assim, cada observação tem dois componentes:

1. Yi-Yi revela a discrepância entre um valor observado e o valor predito correspondente; essa discrepância é o termo de erro (e_i).
2. Yi-Y indica a porção do escore da observação que se deve a regressão linar entre Y e X.

Elevando ao quadrado ambos os lados da identidade acima é somando para todos os N da amostra observada obtem-se a **soma dos quadrados da regressão** e a **soma dos erros ao quadrado**. Se mexermos um pouco nestes termos, temos que:

$$i=1NYi-Y2=i=1NYi-Y2+i=1NYi-Yi2$$

A esquerda temos a **soma total dos quadrados**, ou SQ_{total}. Os dois termos a esquerda representam a soma dos quadrados da regressão e a soma dos quadrados dos erros.

$$SQ_{total} = SQ_{regressão} + SQ_{erro}$$

3.2.1 – O coeficiente de determinação

O coeficiente de determinação da regressão indica a proporção da variação todal de Y que é "determinada" pela relação linear com X. Seu símbolo é R^2_{YX} (ou R ao quadrado) que é obtido pela seguinte formula:

$$Ryx2=Yi-Y2-Yi-Yi2Yi-Y2$$

Ou

$$Ryx2=SQtotal-SQerroSQtotal$$

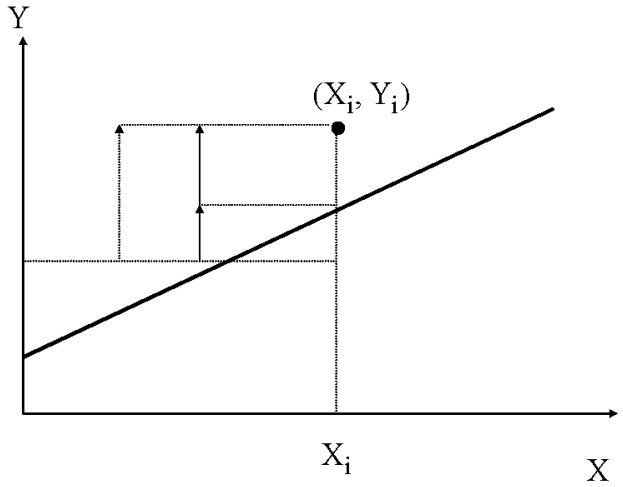
$$Ryx2=1- SQerroSQtotal$$

Assim, o R-quadrado é igual a 1 menos a razão entre a soma do quadrado dos erros e a soma total dos quadrados. Elevar ao quadrado implica em restringir a variação de R^2_{YX} ao intervalo entre 0 e 1. Quando SQ_{erro} for zero $R^2_{YX} = 1$, ou seja, toda variação em Y é explicada pela variável dependente. Quando $SQ_{erro} = SQ_{total}$ (ou seja, toda variação se deve ao erro) $R^2_{YX} = 0$. Uma vez que $SQ_{regressão} = SQ_{total} – SQ_{erro}$, há uma quarta formula para R-quadrado:

$$Ryx2= SQregressãoSQtotal$$

Uma formula simples de ser calculada para a relação bi variada é a que faz a razão entre o quadrado da covariância dividido pelo produto das duas variâncias:

$$Ryx2 = Syx2Sx2Sy2$$



3.2.2 – O coeficiente de correlação:

A raiz quadrada de R^2_{yx}, que resume a relação linear entre duas variáveis continuas, é chamada de **Coeficiente de Correlação de Pearsons** (em homenagem ao estatístico Karl Pearsons). Sua formula é:

$$ryx = Ryx2$$

$$ryx = SyxSySx$$

O coeficiente de correlação é útil porque ele mostra a direção da relação entre X e Y. Um valor positivo ou negativo é atribuído a r_{yx} para indicar a direção da covariância. Este sinal deve ser igual ao sinal do coeficiente da regressão (b_{yx}). O coeficiente de correlação de Pearsons é simétrico como mostra a formula abaixo:

$$ryx = SyxSySx$$

$$ryx = Yi\text{-}YXi\text{-}XN\text{-}1\ SySx$$

2. Precauções na análise da associação

a. Extrapolação: predizer valores de y para valores de x fora da abrangência dos dados

b. Cuidado com pontos influentes (outliers):

c. Correlação não é sinônimo causa (causalidade):
Exemplo crime e educação (agresti)
Número de afogados e consumo de sorvete na Australia (agresti)

d. Paradoxo de Simpson

e. O efeito de variáveis escondidas (não mensuradas) na associação
Sempre há essa possibilidade

f. Variáveis que confundem o efeito (duas independentes associadas)

CAPÍTULO 3 - ANÁLISE DE VARIÂNCIA

O tópico "Análise de Variância" encobre uma gama de modelos para análise de dados, modelos esses que envolvem uma variável dependente *Y* mensurada a nível de intervalo ou razão, e uma ou mais variáveis independentes, escolhidas para explicar as diferenças em *Y*, mensuradas ao nível nominal. Em outras palavras, para cada categoria (ou combinação de categorias) da minha variável independente (ou variáveis independentes), fazemos um conjunto de observações sobre a variável dependente. O propósito da Análise de Variância é descobrir se existem diferenças no nível (médio) da variável dependente *Y* entre as categorias das variáveis independentes. Alternativamente, podemos entender que seu propósito é estimar em que medida o conhecimento da categoria da variável independente/explicativa reduz o erro de predição do valor da variável explicada/dependente Y.

Vamos estar aqui interessados no caso em que apenas uma variável nominal está envolvida, ou seja, na Análise da Variância a Um Fator. Como a variável nominal tem *k* categorias, o nosso problema será, portanto, o de descobrir se esses *k* grupos definidos pelas categorias da variável independente se caracterizam por níveis (isto é, médias) diferentes na variável dependente *Y*.

Assim, os dados para a Análise da Variância podem ser representados na forma

Categoria *i*

i=1	*i*=2	...	*i*=*k*
Y11	Y21	...	Yk1
Y12	Y22	...	Yk2
...	...		...
Y1n1	Y1n2	...	Yknk

Observe-se que para cada um dos *k* grupos temos ni (*i*=1, 2,..,k) observações diferentes . Em outras palavras, o número de observações dentro de cada grupo pode ser diferente, variando de grupo a grupo.

Modelo a 1-Fator

Supõe-se que os dados para a Análise da Variância a 1 fator foram gerados por um modelo que pode ser descrito como

$i = 1, 2, ..., k$
$j = 1, 2, ..., ni$

onde Yij é o valor de *Y* da Jésima observação na iésima categoria. Yij foi produzido pela soma de 3 componentes:

μ = nível geral (média) de *Y*

i= efeito sobre *Y* de pertencer à categoria *i*

eij= efeito sobre *Y* de outras variáveis que não sejam "pertencer a categoria *i*", que se supõe **não relevantes**. Também chamada de "variável residual".

Os verdadeiros valores dos parâmetros μ, 1, 2, ..., k e do termo residual são naturalmente desconhecidos, sua estimação sendo o objeto próprio (bem como o teste de significância de suas eventuais diferenças, o que será discutido na segunda parte de nosso curso) da Análise de Variância.

Para fins de exposição, entretanto, suporemos que para um caso específico em que *k* =3 categorias os parâmetros são os seguintes

$$\mu=10;\ 1=4;\ 2=1;\ 3=-5$$

Observe-se que como μ é o nível (média) geral da variável *Y*, então necessariamente

Conhecidos esses parâmetros e na ausência de efeitos da variável residual, os valores observados seriam (para, por exemplo, categorias com cinco observações cada uma) num dado conjunto de observações

$$Yij=\mu+i$$

Observação	Categoria 1 1=4	Categoria 2 2=1	Categoria 3 3=-5
1	14	11	5
2	14	11	5
3	14	11	5
4	14	11	5
5	14	11	5

$$\therefore\ Y=10$$

A Variação Total entre as 15 observações pode ser medida pela soma dos quadrados

= 210

Como não existe variação dentro de cada grupo (dado que, por enquanto, ignoramos o termo eij), essa Variação Total é completamente originada pelas diferenças entre grupos.

Suponhamos agora que os valores da variável residual eij são os seguintes:

Grupo (categoria *i*)		
1	2	3
-2,5	-3,3	-4
0,1	-3,5	-3,5
3,9	-0,3	0
2,0	0,4	1,0
4,9	2,4	0,4

Assim os valores observados são os seguintes

Grupo (categoria *i*)			
1	2	3	
11,5	7,7	1	Y=9,87
14,1	7,5	1,5	Y1=15,68
17,9	10,7	5	Y2=10,14
16,0	11,4	6	Y3=3,78
18,9	13,4	5,4	

A Variação Total agora é

$$i=1kj=1niYij-Y2=437,38$$

devido ao fato de que, agora, **a Variação Total é uma combinação (soma) da Variação intra-grupos com a Variação inter-grupos.**

Análise de Variância: Estimativas

Os parâmetros μ, α1,2,..., k são estimados por , 1,2,..., k.

A estimativa de Mínimos Quadrados (que será discutida em maior detalhe na próxima aula) implica em minimizar a diferença entre o valor observado e o valor predito pelo modelo.

$$Yij=+i$$

Ou seja, minimiza-se

$$Q=ikjniYij-Yij2=ikjnieij2=ikjniYij--i2$$

sujeito à restrição que

$$iki=0$$

Dessa minimização resultam os seguintes estimadores

e

Assim, para o nosso exemplo, as estimativas dos parâmetros seriam

=Y=9,87

1=Y1-Y=15,68-9,87=5,81

2=Y2-Y=10,14-9,87=0,27

3=Y3-Y=3,78-9,87=-6,09

os quais, naturalmente, discrepam dos valores (desconhecidos) originais.

Decompondo a Variação Total

Observe-se que o desvio para cada observação no conjunto de dados pode ser escrito como

Se agora quadramos esse desvio e somamos ao longo de todas as observações

$$i=1kj=1niYij-Y2=i=1kj=1niYij-Yi+Yi-Y2=$$

$$=i=1kj=1niYij-Yi2+i=1kj=1niYi-Yi2+2i=1kj=1niYij-YiYi-Y$$

Agora, o último termo à direita da equação é

$$2i=1kj=1niYij-YiYi-Y=2i=1kYi-Yj=1niYij-Yi=0=0$$

uma vez que Yi-Y é constante para cada valor de *i* e j=1niYij-Yi=0 necessariamente por definição de Yi

Logo

$$i=1kj=1niYij-Y2=i=1kj=1niYij-Yi2+i=1kj=1niYi-Y2 \text{ ni vezes}$$

Como no último termo $Y_i - \bar{Y}$ é constante para cada valor de *i*, temos que $\sum_{j=1}^{n_i} (Y_i - \bar{Y})^2 = n_i (Y_i - \bar{Y})^2$.
Então

Soma dos Quadrados Total (TSS)	Soma dos Quadrados Intra-Grupos (WSS)	Soma dos Quadrados Inter-Grupos (BSS)

Variância Total	"Variação não-explicada" pela variável categórica	"Variação explicada" pela variável categórica

Ou seja, podemos decompor a Variação Total (Soma dos Quadrados) em 2 componentes: Variação Inter ("Explicada") e Variação Intra-Grupos (Não-Explicada).

Estimando o poder preditivo de X sobre Y

Observe ainda que a "Variação explicada" (BSS) tem uma expressão de simples cálculo, dado que já estimamos os α_i anteriormente:

Variação explicada = BSS = $\sum_{i=1}^{k} n_i (Y_i - \bar{Y})^2 = \sum_{i=1}^{k} n_i \alpha_i^2$

Resultando também ser simples o cálculo da Variação Intra-Grupos (WSS), dado que

$$WSS = TSS - BSS.$$

Efetuando os cálculos para nosso exemplo, obtemos

$BSS = \sum_{i=1}^{k} n_i \alpha_i^2 = 5*(5,81)^2 + 5*(0,27)^2 + 5*(-6,09)^2 = 168,78 + 0,36 + 185,44 = 354,59.$

Como já sabemos que a Variação Total = TSS = 437,38 então resulta que

$WSS = TSS - BSS = 437,38 - 354,59 = 82,79.$

Com esses dados, tentemos agora definir o poder preditivo da variável categórica sobre a variável dependente numérica Y. Lembremos que as medidas que seguem a lógica RPE estão fundadas num conceito preditivo do que é associação. Corresponde, genericamente, a um jogo no qual extrai-se aleatoriamente uma pessoa de nosso conjunto de dados e tenta-se predizer o valor que essa pessoa terá no que diz respeito à variável Y. Inicialmente, esta predição deverá ser feita sem nenhuma informação adicional além da distribuição (univariada/marginal) dessa própria variável Y. Essa é a regra 1 de predição. Em seguida, examina-se o valor de uma segunda variável X e tenta-se predizer o valor de

Y apoiado no conhecimento de X, isto é, na distribuição **condicional** de Y dado o valor de X. Esta é a segunda regra de predição. O critério RPE implica na comparação dos erros de classificação sob essas duas regras. Denotando o erro total de estimação do valor de Y segundo a regra 1 como P(1) e a o erro total de estimação segundo a regra 2 como P(2), a medida de associação que atende o critério RPE tem a forma

$$\text{RPE} = \text{P1-P(2)P(1)}$$

No caso da Análise da Variância, quando só conhecemos a distribuição da variável dependente (explicada) Y, a melhor estimativa que se pode fazer é a sua média Y, sendo a medida de erro total ao se fazer essa predição, naturalmente, a Variação Total (TSS) vista acima:

$$\text{Variação Total} = \text{TSS} = \text{i=1kj=1niYij-Y2.}$$

Ou seja, P(1) = Variação total = TSS = 437,38.

No entanto, quando conhecemos em que categoria i da variável explicativa X o caso selecionado pertence, então a nossa predição não mais será Y, mas agora será ($Yi = \mu + \alpha_i$), sendo que o erro total cometido agora seguindo-se essa segunda regra igual à Variação Intra-Grupos (WSS)

P(2) = Variação Intra-Grupos = Variação "Não-Explicada" = WSS = i=1kj=1niYij-Yi2.

Com isso, podemos definir uma medida de redução proporcional de erro RPE para a análise da Variância com a forma

$$2 = \text{Variação Total-Variação Intra-GruposVariação Total} = \text{i=1kj=1niYij-Y2-} \text{ i=1kj=1niYij-Yi2i=1kj=1niYij-Y2}$$

ou, alternativamente,

$$2 = \text{TSS-WSSTSS} = \text{BSSTSS} = \text{Variação "Explicada" pela variável CategóricaVariação Total}$$

Com isso, 2 **é interpretável tanto como a proporção de redução no erro de predição de Y propiciado pelo conhecimento da variável X, como a proporção da variação de Y que é explicada pela variável X. Como se pode ver, as duas interpretações são "sinônimas".**

No caso de nosso exemplo, que TSS = 437,38; e BSS = 354,59, do que resulta que nossa medida RPE de associação 2 tem o valor

$$2 = 354,59437,38 = 0,81.$$

Ou seja, podemos dizer que o conhecimento da categoria da variável X a qual o nosso caso pertence reduz em 81% o erro de predição do valor de sua variável Y.

Alternativamente, também podemos dizer que as diferenças no pertencimento às categorias de X explicam 81% das diferenças (variância) da variável Y. Esse é a medida do poder explicativo de X sobre Y.

Análise de Variância - Exercícios

1.Suponha que os dados abaixo representam taxas de homicídio entre três tipos de cidades: cidades primariamente industriais, centro comerciais e centros políticos. Para cada grupo fez-se uma amostra de 8 cidades.

Observação	Cidades industriais	Cidades comerciais	Centros políticos	
1	4,3	5,1	12,5	
2	2,8	6,2	3,1	
3	12,3	1,8	1,6	
4	16,3	9,5	6,2	
5	5,9	4,1	3,8	
6	7,7	3,6	7,1	
7	9,1	11,2	11,4	
8	10,2	3,3	1,9	
Σ	68,6	44,8	47,6	Σ=161
$\overline{X}$	8,6	5,6	6,0	$\overline{\overline{X}}$=6,7

Teste a hipótese de que os níveis de homicídios não diferem entre os tipos de cidade.

2.Suponha que os dados abaixos representam salários-hora (em francos) de operários franceses, norte-africanos e portugueses , de uma amostra de operários da indústria automobilística francesa.

Observação	Franceses	Norte-africanos	Portugueses
1	51	40	44
2	47	37	39
3	37	28	33
4	52	53	56
5	42	38	43
6	63	51	56
7	46	45	47
8	62	60	58

Teste a hipótese de que os níveis de salário não diferem entre as nacionalidades dos operários.

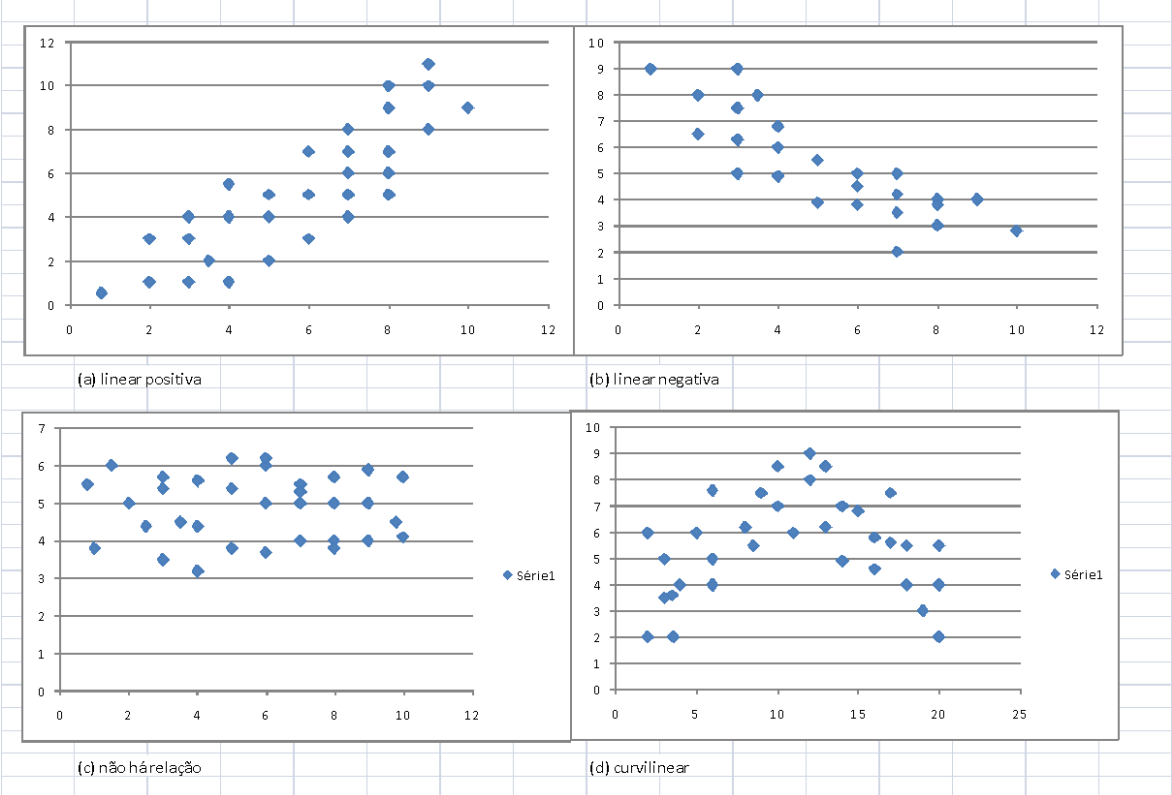
CAPÍTULO 4 – ANÁLISE DE REGRESSÃO

Nesta capítulo apresentamos as técnicas de regressão e correlação para duas variáveis medidas ao nível de intervalo. O caso que apresentamos pressupõe que a forma da relação entre as vaiáveis Y e X é linear e que a variável dependente (Y) se distribui normalmente a cada nível da variável independente (X). Mas mesmo quando os dados violam estas pressuposições o método continua robusto, ou seja, raramente leva a conclusões errôneas.

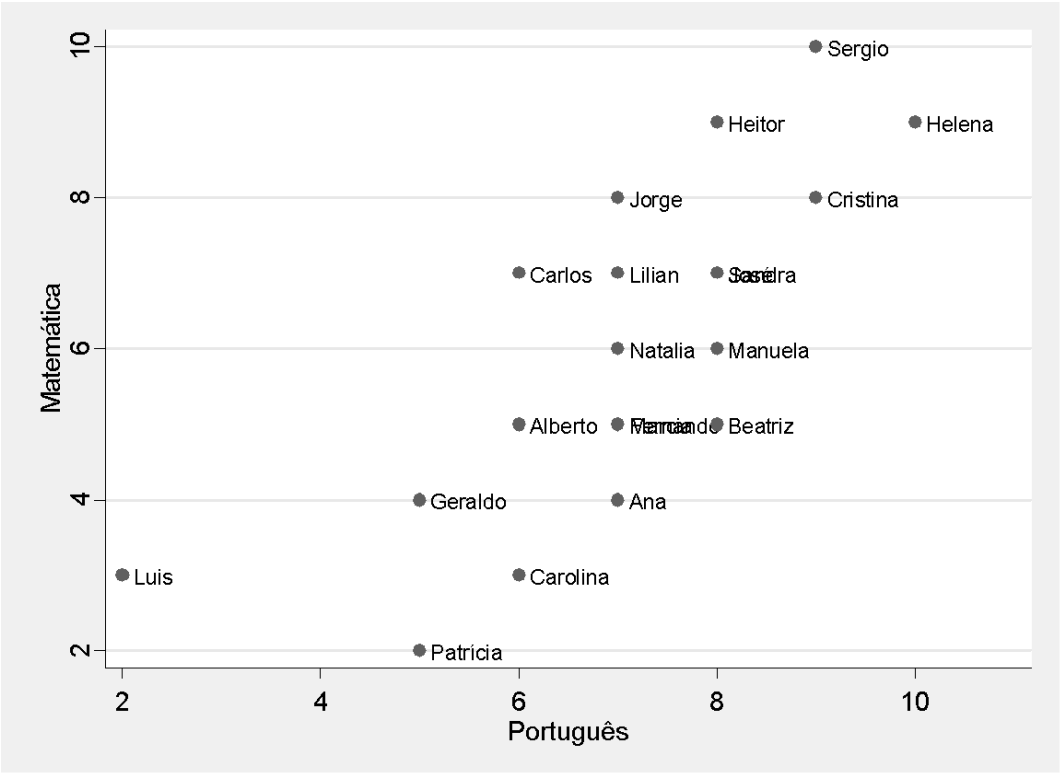
1 - Gráfico de dispersão e Regressão Linear:

Quando trabalhamos com uma relação entre duas ou mais variáveis medidas ao nível de intervalo, a representação gráfica desta relação toma a forma de um eixo de coordenadas cartesianas (ortogonais).

Podemos imaginar quatro possibilidades básicas para a relação entre duas variáveis medidas ao nível de intervalo: (a) relação linear positiva, (b) relação linear negativa, (c) não há relação, e (d) relação curvilinear.



Vamos ver um exemplo mais concreto. Usando o banco de dados hipotético do nosso livro podemos representar a relação entre as notas de Portugês e de Matemática dos vinte indivíduos num gráfico de disperção, onde *x* é a variável no eixo horizontal (Português) e *y* é a variável no eixo vertical (Matemática). Neste caso específico só há valores positivos, mas nada impede que haja valores negativos para *x* e *y* em outros exemplos.



Cada ponto representa o valor de *x* e de *y* para uma dada observação. Como definimos acima, esta representação gráfica é chamada de "diagrama de dispersão". Neste caso temos duas variáveis cuja relação queremos estudar – "nota do indivíduo em matemática" e "nota do indivíduo em Portugês". Os valores observados são os apresentados na tabela abaixo.

caso	Nome	Português	Matemática
1	Alberto	6	5
2	Ana	7	4
3	Beatriz	8	5
4	Carlos	6	7
5	Carolina	6	3
6	Cristina	9	8
7	Fernando	7	5
8	Geraldo	5	4
9	Heitor	8	9
10	Helena	10	9
11	Jorge	7	8
12	José	8	7
13	Lilian	7	7
14	Luis	2	3
15	Manuela	8	6
16	Marcia	7	5
17	Natalia	7	6
18	Patrícia	5	2

19	Sandra	8	7
20	Sergio	9	10

Para estimar a relação linear entre a variável dependente (Y) e a independente (X) fazemos uma "regressão de Y em X", produzindo assim uma relação bi-variada linear, ou de forma simplificada, uma regressão bi-variada. A forma algébrica geral da equação linear bi-variada é a seguinte:

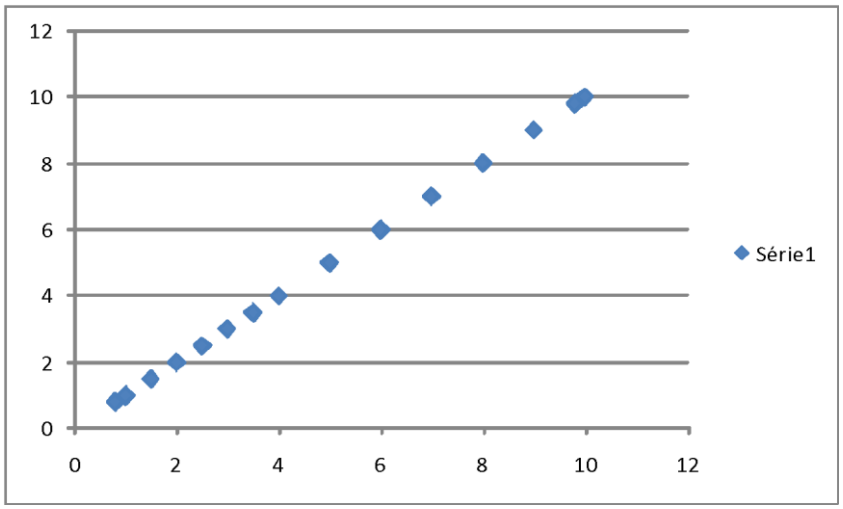
$$Y = a + bX$$

A ordenada, ou valor Y, é igual a soma de uma constante, (a) (o ponto de interseção entre a linha e o eixo Y) mais o produto da inclinação, b, pelo valor de X. A interseção da linha com o eixo Y mostra o valor de Y quando X=0. A inclinação da linha (b) mostra quantas unidades de Y são necessário para que haja uma mudança de uma unidade em X.

Começamos com uma equação de predição em que o valor da i^{esima} observação na variável dependente (Y) é uma função linear exata da sua variável independente (X):

$$Yi=a+byxXi$$

Essa equação prevê uma relação linear perfeita, como por exemplo a representada no gráfico abaixo.



Mas dados sociais reais nunca seguem uma relação linear perfeita, como fica claro no gráfico cruzando "nota em matemática" e "nota em português". Conseqüentemente, ao elaborar o modelo de regressão linear devemos calcular e levar em conta os desvios em relação a predição linear. O modelo de regressão linear segue, portanto, a formula:

$$Yi=a+byxXi+ei$$

Repare que nesta equação não incluímos o acento circunflexo sobre o Y, o que significa que a equação é para o valor observado de Y e não para o predito Y. O termo de erro (e_i) representa a porção do valor da i^{esima} observação que não é predito pela relação linear de Y com X. Logo e_i mede a discrepância que ocorre quando fazemos a predição utilizando a equação da regressão linear. Ao longo de todas as N observações alguns erros de predição serão positivos (>0), outros negativos (<0) e alguns exatamente zero (quando o valor predito for exatamente igual ao observado). Da mesma forma que ocorre com desvios em relação a média (variância), a soma dos erros (Σe_i) é igual zero porque os valores positivos e os negativos se cancelam uns aos outros.

O termo de erro da regressão também é conhecido como resíduo, porque ele é o montante que permanece depois que subtraímos a equação de predição do modelo da regressão linear:

$$Yi-Yi=a+byxXi+ei-a+byxXi= ei$$

A tarefa básica da análise de regressão é estimar os valores para os dois coeficientes da regressão (a e b_{yx}) a partir dos dados observados. As estimativas de a e b_{yx} devem minimizar os resíduos, ou seja, tornar os erros preditos usando a equação menores do que os erros obtidos por qualquer outra relação linear.

2 – Estimando a equação da regressão linear

2.1 – O critério dos mínimos quadrados

Todas as N observações das duas variáveis são usadas para estimar a equação da regressão linear. As estimativas dos coeficientes (a e b_{yx}) estão de acordo com o critério da soma dos mínimos quadrados ordinais. Tendo em vista que a soma dos resíduos (Σe_i) é sempre igual a zero, elevar ao quadrado remove os sinais negativos de forma que a soma dos erros ao quadrado é sempre maior do que zero. Conseqüentemente, a soma das diferenças ao quadrado entre cada observação (Yi) e seu valor predito pela equação da regressão (Yi) produz um valor menor do que o obtido pelo uso de qualquer outra equação linear. Ou seja,

$$i=1NYi+Yi2=i=1Nei2$$

é um mínimo.

Estimativas com esta propriedade de erros mínimos produzem estimadores de **mínimos quadrados ordinais (OLS)** de a e b_{yx}. O método OLS de estimar a linha da regressão é semelhante ao cálculo da média e da variância em torno da média, que apresentamos no capítulo 1. De fato, o método da regressão linear (OLS) estima a média Y condicional em X, ou seja, a média de Y variando de acordo com a variação nos valores de X. De forma semelhanto aos cálculos da média que observamos no capítulo 1, a linha da regressão minimiza a soma dos erros predítos ao quadrado.

O estimador OLS do coeficiente da regressão bi variada, ou inclinação da regressão (b_{yx}), a partir das observações de X e Y é o seguinte:

$$byx=Yi-YXi-XXi+X2$$

O numerador (parte de cima) desta équação é o somatório do produto dos desvios de cada observação Yi e Xi das variáveis X e Y em torno de suas respectivas médias Y e X.

Quando este termo é dividido por N-1, ele é denominado de **covariância**, e frequentemente representado pelo símbolo S_{yx}. Ou seja,

$$Syx=Yi-YXi-XN-1$$

Para calcular a covariância também podemos usar a seguinte equação:

$$Syx=NYiXi-XiYiN(N-1)$$

O denominador (parte de baixo) para a formula de b_{yx} é a soma dos desvios ao quadrado em torno da média da variável independente, X.

Dividindo este termo por N-1 temos a **variância da variável independente (X)**:

$$Sx2=i=1NXi-X2N-1$$

Tendo em vista que N-1 aparece tanto no denominador da covariância quanto na variância da variável dependente (dois quadros acima), podemos cancelar N-1 da razão entre covariância e variância. Logo, o estimador OLS do coeficiente da regressão bi variada pode ser estimado pela seguinte razão:

$$byx=SyxSx2$$

Embora as duas formulas para a e b_{yx} descritas acima sejam conceitualmente precisas, elas são difíceis de serem calculadas sem um computador. De fato, qualquer cientista social que

queira analisar dados usando regressões lineares vai fazer uso de computadores e pacotes estatísticos. No entanto, é didaticamente interessante entendermos o que estes pacotes fazem. Para tanto apresentamos uma fórmula fácil de calcular sem o uso do computados. Essa formula simplificada para calcular b, que não depende de desvios, produz resultados numericamente idênticos e pode ser calculanda usando uma calculadora manual:

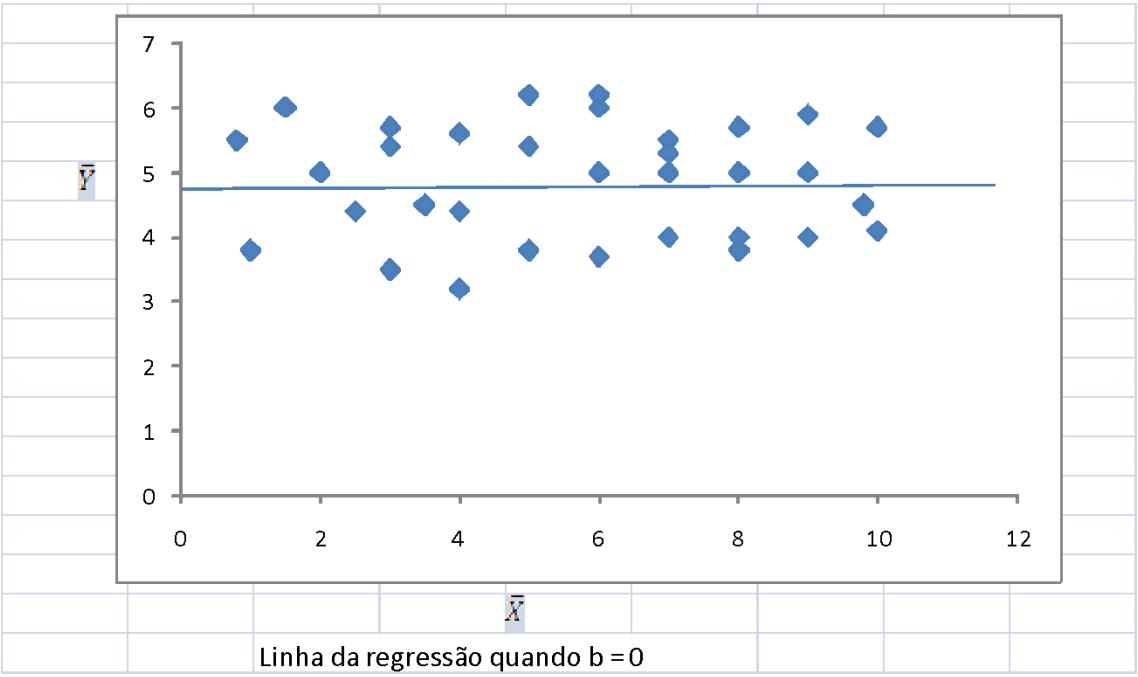
$$byx=NYiXi-YiXiNXi2-Xi2$$

O estimador para o **intercepto** (a) é simplesmente:

$$a=Y-bX$$

Essa equação assegura que o par coordenado (X,Y) sempre fique na linha da regressão, independentemente da valor de a e b. Logo, para calcular o intercepto, o coeficiente da regressão bi variada é calculado antes, e em seguida utilizado na formula acima junto com as médias de ambas as variáveis.

A equação da regressão linear estima uma **média condicional** para Y, ou seja, um valor predito para a variável dependente (Yi) para cada valor especifico da variável independente Xi. Se não houver relação linear entre Y e X, a inclinação (coeficiente da regressão) b_{yx} será igual a zero (como no gráfico abaixo). Ou seja, todos os valores preditos são iguais a a, porque $Yi=a+0Xi$, o que leva a $Yi=a$ e a $Yi=Y$. Quando a inclinação b_{yx} (coeficiente) da regressão é igual a zero, o conhecimento de um valor especifico de Xi não produz um valor predito de Y que seja diferente da media de todas as observações de Y. Neste caso podemos ficar apenas com a média de Y (Y) como nossa melhor opinião sobre a população que estamos estudando. Em outras palavras, se o valor de b_{yx} for igual a zero o conhecimento de X não acrescenta nada ao conhecimento de Y.



Linha da regressão quando b = 0

2.2 – Regressão linear aplicada aos nossos dados sobre nota na prova de matemática e de português

Será que as notas de Português e de Matemática dos 20 alunos de nossa amostra estão correlacionadas. Uma pergunta válida é: será que bons alunos em uma matéria também tem bom desempenho na outra? Se soubéssemos a nota de Português dos alunos teríamos alguma chance de adivinhar a nota de matemática, ou pelo menos de dar um bom palpite. Se ambas as notas estiverem correlacionadas o conhecimento sobre uma delas nos ajuda a entender a outra. De fato, podemos estimar as notas de Matemática usando as de Português, ou seja, se soubermos a nota de Português de algum aluno temos uma boa chance de saber melhor qual sua nota de Matemática. Como vimos no gráfico 2 (diagrama de dispersão das notas) acima quanto maior a nota de Português do aluno maior sua nota de Matemática. Vejamos como analisar estes dados usando uma regressão linear estimada pelo método de Mínimos Quadrados Simples (OLS). A tabela abaixo apresenta as notas em ambas as matérias e os principais cálculos que precisamos fazer para estimar os parâmetros a e b_{yx} da regressão com o objetivo de definir qual a linha da regressão.

Tabela 1

nome	n	Matemática Y	Português X	(Xi - X)	(Yi - Y)	(Xi - X)(Yi - Y) =	$(Xi - X)^2$	$(Yi - Y)^2$	a	byx	Y^	(Yi-Y^)2	(Yi-Ymed)2	(Y^-Ymed)2	YX
Alberto	1	5	6	-1,0	-1,0	1,0	1,0	1,0	-0,638	0,948	5,052	0,003	1,000	0,899	30
Ana	2	4	7	0,0	-2,0	0,0	0,0	4,0	-0,638	0,948	6,000	4,000	4,000	0,000	28
Beatriz	3	5	8	1,0	-1,0	-1,0	1,0	1,0	-0,638	0,948	6,948	3,796	1,000	0,899	40
Carlos	4	7	6	-1,0	1,0	-1,0	1,0	1,0	-0,638	0,948	5,052	3,796	1,000	0,899	42
Carolina	5	3	6	-1,0	-3,0	3,0	1,0	9,0	-0,638	0,948	5,052	4,210	9,000	0,899	18

Cristina	6	8	9	-2,00	-2,00	4,0	4,0		-0,638	0,948	7,897	0,011	4,000	3,597	72
Fernando	7	5	7	0,00	-1,00	0,0	1,0		-0,638	0,948	6,000	1,000	1,000	0,000	35
Geraldo	8	4	5	-2,00	-2,00	4,0	4,0		-0,638	0,948	4,103	0,011	4,000	3,597	20
Heitor	9	9	8	1,00	3,00	1,0	9,0		-0,638	0,948	6,948	4,210	9,000	0,899	72
Helena	10	9	10	3,00	3,00	9,0	9,0		-0,638	0,948	8,845	0,024	9,000	8,093	90
Jorge	11	8	7	0,00	2,00	0,0	4,0		-0,638	0,948	6,000	4,000	4,000	0,000	56
José	12	7	8	1,00	1,00	1,0	1,0		-0,638	0,948	6,948	0,003	1,000	0,899	56
Lilian	13	7	7	0,00	1,00	0,0	1,0		-0,638	0,948	6,000	1,000	1,000	0,000	49
Luis	14	3	2	-5,00	-3,00	15,0	25,0	9,0	-0,638	0,948	1,259	3,032	9,000	22,481	6
Manuela	15	6	8	1,00	0,00	0,0	1,0	0,0	-0,638	0,948	6,948	0,899	0,000	0,899	48
Marcia	16	5	7	0,00	-1,00	0,0	0,0	1,0	-0,638	0,948	6,000	1,000	1,000	0,000	35
Natalia	17	6	7	0,00	0,00	0,0	0,0	0,0	-0,638	0,948	6,000	0,000	0,000	0,000	42
Patrícia	18	2	5	-2,00	-4,00	8,0	4,0	16,0	-0,638	0,948	4,103	4,424	16,000	3,597	10
Sandra	19	7	8	1,00	1,00	1,0	1,0	1,0	-0,638	0,948	6,948	0,003	1,000	0,899	56
Sergio	20	10	9	2,00	4,00	8,0	4,0	16,0	-0,638	0,948	7,897	4,424	16,000	3,597	90
somatório		120	140	0,00	0,00		58,0	92,0				39,8	92,0	52,2	895
Média		6,000	7,000												
Variância							3,053	4,842							
Des. Padrão (raiz da variância)							1,7472	2,2005							

Usando os dados da tabela acima e a fórmula abaixo (poderíamos usar a fórmula alternativa e chegaríamos ao mesmo resultado) obtermos o seguinte valor:

$$byx=Yi-YXi-XXi+X2$$

$$byx=0,948276$$

Uma mudança de um ponto na nota de português corresponde à uma mudança de 0,94 pontos na nota de Matemática. Uma vez que a média de X (da prova de português) é 7, a média de Y (prova de matemática) é 6 e $byx=0,948276$ podemos facilmente calcular o valor de a (intersseção):

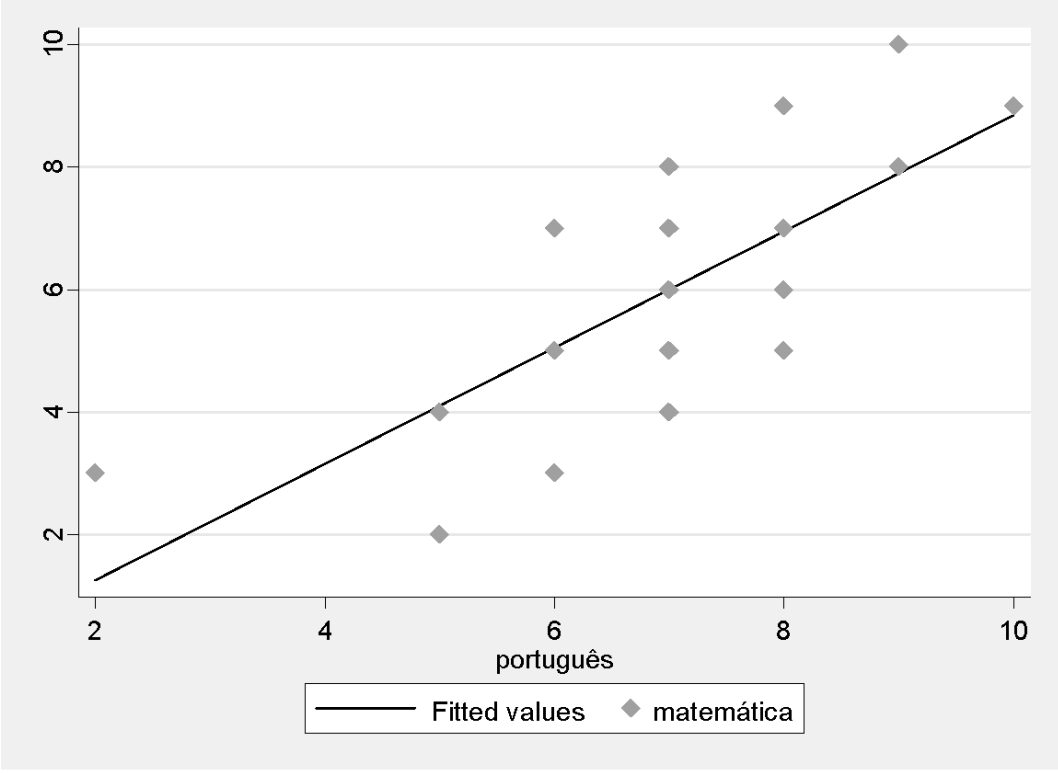
$$a=Y-bX$$

$$a=-0,63793$$

Assim a equação da regressão linear por mínimos quadrados para os dados sobre as notas de Portugês e Matemática é:

$$Yi=-0,63793+0,948276Xi$$

Esta equação nos permite traçar a linha de regressão linear para os nossos dados. O gráfico abaixo apresenta essa reta estimada pelo métodos os mínimos quadrados apresentado acima.



O gráfico acima mostra que a linha da regressão é o melhor estimador da relação entre nota de Português e nota de Matemática, mas não é um estimador perfeito, ou seja, há pontos observados que ficam acima e outros que ficam abaixo da reta estimada. Estas diferenças em torno da linha estimada (da média Y condicional em X) constituem o erro de estimativa. Quanto mais distantes da reta forem os pontos observados maior será o erro, e quanto menos

distantes menor será o erro. Obviamente queremos um erro que seja o menor possível para os dados em questão, e é justamente isso que o métodos apresentado nesta seção faz. De qualquer forma, dados sobre o mundo social nunca se conformam exatamente a nossas estimativas. Por este motivo também devemos ter uma medida do tamanho do nosso erro. Reparem que em qualquer pesquisa empírica temos erros em alguma medida. Nunca sabemos com certeza absoluta se nossas explicações e estimativas sobre o que estamos observando são totalmente corretas. A análise quantitativa de dados em ciências sociais nos permite estimar o tamanho desse erro.

3 – R^2 e coeficiente de correlação

Uma maneira de determinar qual a força da covariação entre duas variáveis é medir a distância média entre os valores observados e a linha da regressão. Em qualquer regressão usada nas ciências sociais há alguma quantidade de erro e, portanto, o objetivo da análise de dados usando a regressão é determinar a contribuição relativa da predição e do erro para explicar a variação que observamos na veiável dependente. Em outras palavras, a variação de Y deve-se tanto ao efeito de X quando ao erro aleatório, o que significa que podemos partir a soma total dos quadrados em uma distribuição de escores entre um componente sistemático e outro aleatório.

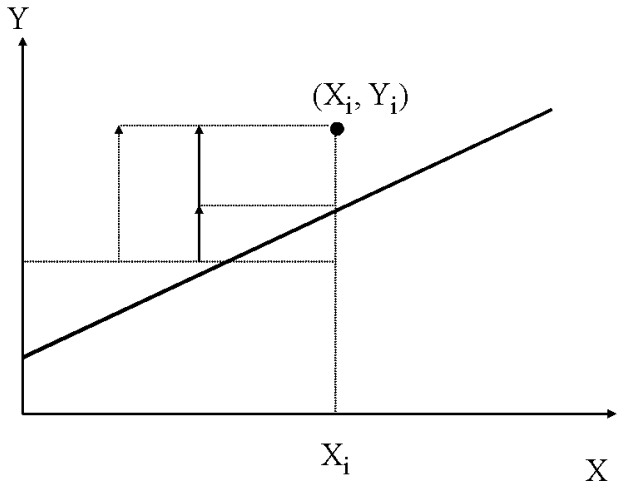
Inicialmente crie um desvio subtraindo a média $\bar{Y}$ do valor observado Y_i. Em seguida adicione e diminua o valor predito pela regressão linear ($\hat{Y}i$) deste desvio, produzindo assim a seguinte identidade:

$$Yi-\bar{Y}=\hat{Y}i-\bar{Y}+Yi-\hat{Y}i$$

Assim, cada observação pode ser vista como tendo dois componentes.

1. $\hat{Y}i-\bar{Y}$ indica a porção do escore da observação que se deve a relação linar entre Y e X. O ponto de comparação é a média de Y, ou seja, o valor médio de Y quando não levamos em conta X.
2. $Yi-\hat{Y}i$ revela a discrepância entre um valor observado e o valor predito correspondente; essa discrepância é o termo de erro (e_i).

Esta idéia é facilmente observada no gráfico abaixo.



Mas como sabemos precisamos dessas informações não apenas para uma única observação, mas sim para todas as observações do nosso banco de dados. Para tanto devemos somar todas as diferenças o que, como também já observamos, levaria a um valor igual a zero. Para que as diferenças positivas e negativas não se anulem seguimos o mesmo procedimento que usamos para o cálculo da variância: somamos as diferenças de todos os valores ao quadrado. Elevando ao quadrado ambos os lados da identidade acima e somando para todos os N valores da amostra observada obtem-se a **soma dos quadrados da regressão** e a **soma dos erros ao quadrado**. Se mexermos um pouco nestes termos, temos que:

$$i=1NYi-Y2=i=1NYi-Yi2+i=1NYi-Y2$$

A esquerda temos a **soma dos quadrados total**, ou **SQTotal (SST)**. Os dois termos a esquerda representam a **soma dos quadrados da regressão**, **SQRegressão (RSS)**, e a **soma dos quadrados dos erros, SQErros (ESS).**

SQTotal = SQRegressão + SQErros

De acordo com esta equação, assumindo que Y e X estão linearmente correlacionadas, podemos dizer que toda a variação observada na variável dependente Y (SQTotal) pode ser dividida em duas partes: uma relacionada a sua relação linear com a variável (ou variáveis) independente (SQRegressão), e outra devida à erros de predição (SQErro).

4 – O coeficiente de determinação

O coeficiente de determinação da regressão indica a proporção da variação total de Y que é "determinada" pela relação linear com X. Seu símbolo é R^2_{yx} (ou R ao quadrado) que é obtido pela seguinte formula:

$$Ryx2=i=1NYi-Y2-i=1NYi-Yi2i=1NYi-Y2$$

ou

$$Ryx2=SQTotal-SQErroSQTotal$$

$$Ryx2=1- SQErroSQTotal$$

Assim, o R-quadrado é igual a 1 menos a razão entre a soma do quadrado dos erros e a soma total dos quadrados. Elevar ao quadrado implica em restringir a variação de R^2_{yx} ao intervalo entre 0 e 1. Quando SQErro for zero $R^2_{yx} = 1$, ou seja, toda variação em Y é explicada pela variável dependente. Quando SQErro= SQTotal (ou seja, toda variação se deve ao erro) $R^2_{yx} = 0$. Uma vez que SQRegressão = SQTotal – SQErro, há uma quarta formula para R-quadrado:

$$Ryx2= SQRegressãoSQTotal$$

Uma fórmula simples de ser calculada para a relação bi variada é a que faz a razão entre o quadrado da covariância dividido pelo produto das duas variâncias:

$$Ryx2= Syx2Sx2Sy2$$

Aplicando esta formula para os dados sobre as notas em Matemática e Português, temos que:

$$Ryx2= SQRegressãoSQTotal$$

$$Ryx2= 52,292=0,5669$$

Alternativamente poderíamos ter usado a última fórmula e obteríamos:

$$Ryx2= Syx2Sx2Sy2$$

$$Syx=NYiXi-XiYiN(N-1)$$

$$Syx=20895- (120)(140)20(20-1)=2,894737$$

Usando a coluna das somas da Tabela 1 podemos calcular $Sx2=3,053$ e $Sy2=4,842$. Assim, o R^2 fica:

$$Ryx2= (2,894737)23,053(4,842)=0,5669$$

Podemos dizer que 56,7% da variação na nota de matemática é “explicada” pela nota em português. Em outras palavras poderíamos dizer que o conhecimento da nota de português elimina 56,7% do erro de predição sobre a nota de matemática. Ou seja, se não soubéssemos como as duas notas estão relacionadas, nossa melhor predição sobre a nota de matemática para cada aluno (Y_i) seria simplesmente a média da nota de matemática (Y). O conhecimento do valor de X, no caso da nota em português, elimina 56,7% do erro que cometeríamos estimando a nota de matemática de cada aluno sem sem saber sua nota de português. A lógica apresentada nas fórmulas acima permitem calcular a Redução

Proporcional do Erro (RPE), que como vimos anteriormente é uma das nossas medidas de associação entre duas ou mais variáveis.

4 – O coeficiente de correlação:

A raiz quadrada de R^2_{yx}, que resume a relação linear entre duas variáveis continuas, é chamada de **Coeficiente de Correlação de Pearsons** (em homenagem ao estatístico Karl Pearson). Sua formula é:

$$ryx = Ryx2$$

$$ryx = SyxSySx$$

O coeficiente de correlação é útil porque ele mostra a direção da relação entre X e Y. Um valor positivo ou negativo é atribuído a r_{yx} para indicar a direção da covariância. Este sinal deve ser igual ao sinal do coeficiente da regressão (b_{yx}). Para o exemplo das notas de matemática e português o r_{yx} = +0,75293 (este tipo de coeficiente varia entre –1 e +1). O coeficiente de correlação de Pearson é simétrico, ou seja, o coeficiente de correlação entre X e Y é igual ao coeficiente de correlação entre Y e X.

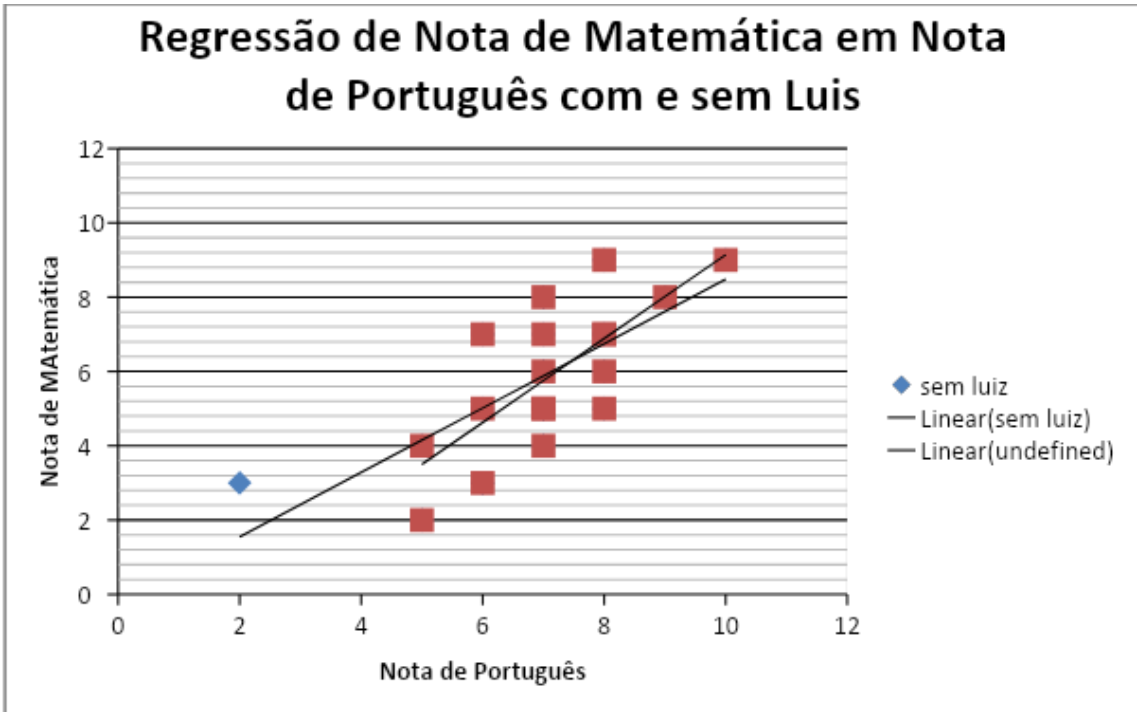
(talvez incluir relação entre coeficiente de correlação e coeficiente da regressão aqui – Treiman pp. 94).

5 – Alguns cuidados na análise de regressão e da associação.

5.1 – Pontos influentes e pontos fora da curva (outliers)

As estatísticas de coeficiente de correlação e de regressão são muito sensíveis à observações que se desviem muito dos padrões típicos da distribuição. Essa sensibilidade está diretamente ligada ao critério dos mínimos quadrados utilizado, ou seja, pelo fato de os "erros" (diferença entre os valores preditos e os observados na variável dependente) serem elevados ao quadrado quanto maior for o erro mais ele vai contribuir para a soma dos erros ao quadrado. Consequentemente, o coeficiente da regressão (b) pode ser afetado de forma substancial por uma ou algumas poucas observações desviantes, que podem puxar fortemente a linha da regressão em sua direção produzindo resultados enganosos.

Para observar isso podemos olhar mais uma vez para os dados sobre notas de matemática e de português.



Regressão incluindo todos os casos

Source	SS	df	MS
Model	52,2	1	52,2
Residual	39,8	18	2,2
Total	92	19	4,8

R-squared	=	0,5669
Adj R-squared	=	0,5428
Root MSE	=	1,4878

matem_ti	Coef,	Std, Err,
portugu_	0,95	0,20
_cons	-0,64	1,41

Regressão excluindo Luis (aluno com notas muito abaixo dos outros)

Source	SS	df	MS
Model	48,5	1,0	48,5
Residual	34,0	17,0	2,0
Total	82,5	18,0	4,6

R-squared	=	0,588
Adj R-squared	=	0,5638
Root MSE	=	1,4142

matem_ti	Coef,	Std, Err,
portugu_	1,24	0,25
_cons	-2,83	1,85

Outros problemas:

Truncation (Treiman)

Regression towards the mean (Treiman)

Extrapolation (Agresti)

Correlation does not imply causation (and Simpson´s Paradox) (Agresti)

The effect of lurking variables on association (Agresti)

Confounding (Agresti)

PARTE 2: PROBABILIDADE E INFERÊNCIA

CAPÍTULO 5 – PRINCÍPIOS BÁSICOS DE PROBABILIDADE

Probabilidade

I – Probabilidade

I. 1. Conjuntos e o Espaço Amostral

Não estamos aqui interessados no cálculo de probabilidades propriamente dito, mas apenas nos resultados mais gerais da teoria de probabilidades. Para desenvolver as noções de probabilidade em que estamos interessados utilizaremos a chamada teoria de conjuntos.

Um conjunto é uma coleção de objetos bem definidos e distintos. Por exemplo, os números 1, 2 e 3 podem formar um conjunto, escrevendo-se

$$S = \{1,2,3\}$$

Observe-se que o conjunto para os cinco números 1,2,2,3,3 é também o conjunto descrito acima.

Um conjunto vazio ou nulo é um conjunto que não contém nenhum elemento, sendo escrito

$$S = \emptyset$$

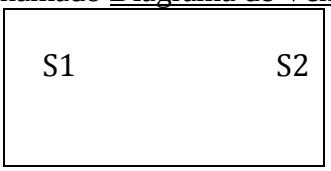
Se todos os elementos pertencentes ao conjunto S1 pertencem também ao conjunto S, então diz-se que S1 é um subconjunto de S e escreve-se

$$S1 \subset S$$

União de conjuntos S1 e S2 é definido como o conjunto dos elementos que pertencem à S1ou à S2 ou a ambos. Se S é a união dos conjuntos S1 e S2 escrevemos

S= S1S2

Diagramaticamente, usando o chamado Diagrama de Venn

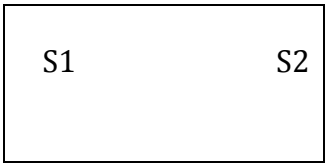


S= S1S2

Interseção de Conjuntos S1 e S2 é o conjunto dos elementos que pertencem simultaneamente à S1e à S2 . Se S é a interseção de S1 e S2 escrevemos

S= S1S2

Diagramaticamente



S= S1S2

Exemplo: Suponhamos

S1={1,2,3,4} e S2={1,3,4,5}

então

(a) S= S1S2={1,2,3,4,5}

(b) S= S1S2={1,3,4}

Em probabilidades o conjunto mais importante é chamado espaço de resultados ou espaço amostral: é definido como o conjunto cujos elementos (notados ei) representam todos os resultados bem distintos de um experimento natural ou elaborado artificialmente. Por exemplo o espaço amostral correspondendo ao experimento de se jogar uma moeda tem dois resultados possíveis

S={e1= cara, e2= coroa} ;

O espaço amostral de jogar um dado tem seis elementos

S={e1=1, e2=2, e3=3 , e4=4, e5=5, e6=6}

Similarmente o espaço amostral do experimento de jogar três moedas (ou jogar uma moeda três vezes) tem oito elementos

S=e1=cara, cara, cara e2=cara, cara, coroa e3=cara, coroa, cara e4=cara, coroa, coroa e5=coroa, cara, cara e6=coroa, cara, coroa e7=coroa, coroa, cara e8=coroa, coroa, coroa

Um espaço amostral que consiste em um número finito de elementos (ou um número infinito mas com elementos que podem ser contados) é chamado de espaço amostral discreto. Qualquer outro espaço amostral é chamado contínuo.

I.2. Axiomas Básicos da Teoria da Probabilidade

Sejam e1, e2, e3 elementos de um espaço amostral S e sejam Pe1, Pe2, Pe3 suas respectivas probabilidades. Postula-se que

Axioma 1 - para qualquer $e1 \subset S$

Axioma 2 –

Com estes axiomas e mais as operações vistas na teoria de conjuntos podemos desenvolver uma teoria matemática extremamente poderosa chamada Teoria das Probabilidades.

I.3. Eventos

Um evento é um subconjunto do espaço amostral. Por exemplo, no espaço amostral do experimento jogar duas moedas

S=cara, cara=e1 cara, coroa=e2 coroa, cara= e3 coroa, coroa=e4

Podemos definir o evento E

E: pelo menos uma cara

Que inclui os resultados e1, e2, e3

E= e1, e2, e3

Relacionado com a definição de evento, temos o seguinte axioma

Axioma 3 –

onde o índice i define os elementos no evento S1

Assim, no exemplo acima

$$PE = Pe1 + Pe2 + Pe3 = 14 + 14 + 14 = 34$$

Como vemos, um evento é um subconjunto qualquer do espaço de resultados (amostral), arbitrariamente definido.

I.4. Teoremas Básicos da Teoria de Probabilidades

Teorema 1 – Se S1 é um evento significando 'não-S1', então

Decorre diretamente dos axiomas básicos. Diagramaticamente

Exemplo: no exemplo acima E= pelo menos 1 cara. Logo E= nenhuma cara.
Então

$$PE1 = 1 - PE = 1 - 34 = 14$$

Teorema 2 (Teorema da Adição) – Sejam dois eventos S1 e S2. Então

Seguem dos axiomas básicos e das operações dos conjuntos. Diagramaticamente

PS1S2 é subtraído para compensar a dupla contagem.

Exemplo: Sejam os eventos E : pelo menos uma cara e G: ambos os resultados iguais do experimento jogar 2 moedas. Como já vimos

$$E=e1, e2, e3 \quad e \quad PE=34$$

Agora, $G=e1, e4 \quad e \quad PG=24$

No entanto, e1 pertence a ambos os eventos:

$$E \cap G=e1 \quad e \quad PE \cap G=14$$

Logo:

$$PE \cup G=PE+PG-PE \cap G$$

$$=34+24-14=3+2-14=44=1$$

Ou seja, $PE \cup G$ cobre todo o espaço amostral.

Se S1 e S2 são mutuamente exclusivos, ou seja $S1S2=\emptyset$, então

$$PS1S2=PS1+PS2$$

Diagramaticamente

Exemplo: no experimento de jogar duas moedas, podemos definir

G: exatamente 1 cara $\therefore PG=24$
H: duas coroas $\therefore PH=14$
$PG \cup H=PG+PH=24+14=34$

Como para cada evento S1 podemos sempre definir seu complemento S1, é bastante útil a representação do espaço amostral em termos de uma tabela cruzada relacionando dois elementos S1 e S2. Ou seja, podemos construir uma tabela em que as probabilidades de S1 e S2 estão representadas com a forma

	S2	S2	
S1	PS1S2	PS1S2	PS1
S1	PS1S2	PS1S2	PS1
	PS2	PS2	PS=1

Por exemplo, suponhamos que em um grupo de 100 pessoas encontramos 60 fumantes e 55 homens. Se definirmos o experimento "selecionar ao acaso um indivíduo do grupo", podemos ter os eventos

$$S1=Homem$$

$$S2=Fumante$$

E se obtivermos todas as probabilidades destes eventos combinados entre si e com os seus complementos, podemos chegar a uma tabela hipoteticamente com a seguinte forma

		Hábito		
		S2=Fumante	S2 = Não fumante	
Sexo	S1=Homem	40100	15100	55100
	S1=Mulher	20100	25100	45100
		60100	40100	1

Assim, a probabilidade de se obter 'homem' ou 'fumante' é dado pela adição

$$PS1S2=PS1+PS2\ -PS1S2=$$

$$=0{,}55+0{,}60-0{,}40=0{,}75$$

Observe-se que também

$$PS1S2=1-PS1S2=1-0{,}25=0{,}75$$

As probabilidades nas células da tabela (ou seja, as interseções) são chamadas 'probabilidades conjuntas'; as probabilidades nas beiras da tabela são chamadas 'probabilidades marginais', representando as probabilidades singulares de cada evento.

Teorema 3 (Probabilidade Condicional) – Se S1 e S2 são eventos em um espaço amostral discreto e $PS1\neq0$, então a probabilidade de S2 dado S1 é

A probabilidade condicional implica em verificar a probabilidade de S2 ocorrer dada a restrição (condição) de que S1 também ocorra. Por exemplo, a probabilidade de se obter um fumante dado que o indivíduo selecionado é do sexo masculino é

$$PS2|S1=0{,}400{,}55\approx0{,}73$$

É importante que se observe que a probabilidade condicional inversa é diferente daquela acima:

$$PS1|S2=PS1S2PS2$$

Assim, dado que geralmente $PS1\neq PS2$, então geralmente

$$PS1|S2\neq PS2|S1$$

Num exemplo numérico usando o exemplo anterior

$$PS1|S2 = 0{,}40 \, 0{,}60 \approx 0{,}67$$

Teorema 4 (Independência) – Se $PS1 \neq 0$ e $PS2 \neq 0$, então S1 e S2 são independentes se e apenas se

Ou seja, a definição de independência acima indica que S1 e S2 são ditos independentes se a probabilidade condicional de S1 dado S2 for igual à probabilidade não-condicional de S1. A condição de S2 ocorrer não altera a probabilidade de S1 ocorrer.
É importante que se observe que essa definição de independência implica em que

$$PS1S2 = PS1 \times PS2$$

Isso por que, segundo o teorema 3

$$PS1|S2 = PS1S2PS2$$

Como a definição de independência é $PS1|S2 = PS1$, substituindo-se o termo à esquerda na expressão acima temos

$$PS1 = PS1S2PS2$$

$$\therefore PS1S2 = PS1 \times PS2$$

Assim, para o caso hipotético de independência e dadas as distribuições marginais de S1 e S2 , podemos calcular a distribuição conjunta de S1 e S2 no caso daquela hipótese ser verdadeira, utilizando a fórmula acima

	S2	S2	
S1	0,33	0,22	0,55
S1	0,27	0,18	0,45
	0,60	0,40	

Assim, $PS1S2 = 0{,}55 \, 0{,}60 = 0{,}33$ etc.

A comparação dessa tabela hipotética para o caso de independência com a tabela "real" anteriormente fornecida indica que os eventos 'fumante' e 'homem' não são independentes, uma vez que as distribuições conjuntas desses eventos são diferentes.

CAPÍTULO 6 – A DISTRIBUIÇÃO BINOMIAL

Variáveis Aleatórias Discretas e a Distribuição Binomial

I. 1. Variáveis Aleatórias Discretas
Consideremos o espaço amostral correspondendo ao experimento de jogar duas moedas:

$$S=\text{coroa, coroa, coroa, cara,carac, coroa, cara, cara}$$

Podemos associar a cada um dos elementos do espaço amostral um número de caras (ou de coroas), de tal forma que pudéssemos compor a tabela:

Elementos do espaço	Número de caras
Coroa, coroa	0
Coroa, cara	1
Cara, coroa	1
Cara, cara	2

O princípio de se associar uma característica numérica ao espaço amostral pode obviamente ser aplicado a qualquer espaço amostral e a essa característica numérica dá-se o nome de variável aleatória (ou estocástica) discreta. Assim,

uma variável aleatória discreta é uma variável cujos valores estão associados aos elementos de um espaço amostral

Notação: Normalmente se denota a variável aleatória através de uma letra maiúscula (ex: X) e seus respectivos valores através de letras minúsculas (ex: x), freqüentemente acompanhado de um subscrito (ex: xi).

A cada elemento do espaço amostral está associada uma certa probabilidade. Daí decorre que, como cada valor da variável aleatória está associada a um ou mais elementos do espaço amostral, podemos associar uma probabilidade a cada valor da variável aleatória. Por exemplo, no experimento de se jogar duas moedas cada elemento do espaço amostral é composto de dois eventos **independentes**.

Pelo teorema da independência, podemos calcular na hipótese de que $P\text{cara}=P\text{coroa}=12$:

$$P\text{coroa, coroa}=P\text{coroa}\times P\text{coroa}=1212=14$$

$$P\text{coroa, cara}=P\text{coroa}\times P\text{cara}=14$$

$$P_{cara,\ coroa} = P_{cara} \times P_{coroa} = \frac{1}{4}$$

$$P_{cara,\ cara} = P_{cara} \times P_{cara} = \frac{1}{4}$$

Com estes elementos podemos construir a seguinte tabela:

Elementos do espaço	Número de caras: x	Probabilidade de x:px
Coroa, coroa	0	$\frac{1}{4}$
Coroa, cara	1	$\frac{1}{4}$
Cara, coroa	1	$\frac{1}{4}$
Cara, cara	2	$\frac{1}{4}$
$\sum$	-	1

Alternativamente a informação acima fica mais sucintamente representada pela tabela

Número de caras: x	Probabilidade de x:px
0	$\frac{1}{4}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$
$\sum$	1

A tabela acima é chamada de <u>função de probabilidade</u>.
Generalizando:

Se X é uma variável aleatória discreta em valores x1,x2,..., xn e com as probabilidades associadas px1, px2,...pxn, então o conjunto de pares

x1 px1
x2 px2

xn pxn

é chamado de <u>função</u> (ou <u>distribuição</u>) <u>de probabilidades de X</u>.

Um exemplo de função de probabilidade: seja o experimento jogar 2 dados. Sobre o espaço amostral correspondente a esse experimento podemos definir a variável aleatória "soma dos pontos obtidos", com valores variando naturalmente entre 2 e 12. Na hipótese de que os eventos são equiprováveis e independentes, podemos construir a seguinte distribuição de probabilidades:

x	Elementos	px
2	(1,1)	$\frac{1}{36}$
3	(1,2) (2,1)	$\frac{2}{36}$
4	(1,3) (3,1) (2,2)	$\frac{3}{36}$
5	(1,4) (4,1) (2,3) (3,2)	$\frac{4}{36}$
6	(1,5) (5,1) (2,4) (4,2) (3,3)	$\frac{5}{36}$
7	(1,6) (6,1) (2,5) (5,2) (3,4) (4,3)	$\frac{6}{36}$
8	(2,6) (6,2) (3,5) (5,3) (4,4)	$\frac{5}{36}$
9	(3,6) (6,3) (4,5) (5,4)	$\frac{4}{36}$
10	(4,6) (6,4) (5,5)	$\frac{3}{36}$
11	(5,6) (6,5)	$\frac{2}{36}$
12	(6,6)	$\frac{1}{36}$
$\sum$		1

I.2 – Esperança Matemática

Distribuições aleatórias, como toda e qualquer distribuição, possuem características que descrevem sua forma. As características mais essenciais (como já vimos) são a média e a variância, as quais são definidas através dos chamados valores esperados ou esperanças (expectancias) matemáticas.

O Valor Esperado ou Esperança da variável aleatória X é definido como

E, como já vimos, essa é a expressão para a média da população.
Assim

O termo valor esperado é usado para sublinhar o fato que o resultado que se pode antecipar de um experimento é o da média da variável aleatória associada a um experimento, ou seja, a me dia da população X.
Por exemplo o valor esperado para a variável aleatória 'número de pontos' no experimento 'jogo de dois dados' é (nas hipóteses vistas anteriormente)

$$EX=2(136)+3236+4336+...+9436+10336+11(236)+12136=$$

$$=2+6+12+20+30+42+40+36+30+22+1236=25236=7$$

$$EX=X=7$$

O cálculo das expectancias pode ser também aplicado a funções da variável X, como por exemplo, a transformações lineares de X. Daí o seguinte teorema

Teorema 5 – Se X é uma variável aleatória e a e b são constantes, então

$$EaX+b=a\ EX+b$$

Prova:

$$EaX+b=axi+bpxi$$

$$=axipxi+b\ pxi$$

$$=axipxi\_EX+b\ pxi\_1$$

$$=a\ EX+b$$

Como podemos ver as regras aplicáveis ao cálculo de somatórios podem ser diretamente aplicáveis ao cálculo de expectancias.

A variância da variável aleatória X também é definida em termos de expectancia

Uma maneira alternativa de se determinar a variância é

$$X2=VarX=EX-\mu 2=EX2-2X\mu+2$$

$$=EX2-2\mu EX\_+E2$$

$$=EX2-22+2$$

$$=EX2-2$$

Temos ainda o seguinte teorema, relacionado com uma desmembração semelhante vista quando examinamos transformações lineares.

Teorema 6 – Se X é uma variável aleatória e a e b são constantes, então

$$VaraX+b=a2\ VarX$$

Prova:

$$VaraX+b=E[aX+b-EaX+b\_a\mu+b]2$$

$$=E[aX+b-a\mu+b]2$$

$$=E(aX+b-a\mu+b)2=E(aX-a\mu)2$$

$$=E[aX-\mu]2=E[a2(X-\mu)2]$$

$$=a2E(X-\mu)2\_VarX$$

$$VaraX+b=a2\ VarX$$

I.3 – A Distribuição binomial

Um dos tipos mais importantes de distribuição de variável aleatória é a chamada distribuição binomial. Uma variável aleatória tem uma distribuição binomial quando o experimento a ela associado é composto de n 'ensaios' independentes, o resultado de cada um desses ensaios podendo ser "sucesso" ou "fracasso" com probabilidades respectivamente π e $(1-\pi)$.

Exemplo: Experimento: gerar três filhos
Variável X = número de filhos homens
$\pi=12$ e $1-\pi=12$

Consideremos o espaço amostral associado a essa variável X (denotando H para filho homem e M para as filhas):

Espaço amostral	Valor de x
MMM	0
HMM	1
MHM	1
MMH	1
HHM	2
HMH	2
MHH	2
HHH	3

A probabilidade de se obter uma seqüência particular é dada pelo produto das probabilidades de cada resultado naquela seqüência. Por exemplo, a probabilidade da seqüência

$$HMH = 1\text{-}\pi\pi{=}2(1\text{-}\pi)$$

Ou seja, a probabilidade de se obter x "sucessos" em n ensaios em uma seqüência específica é dada por

$$x\ 1\text{-}\pi n\text{-}x$$

Para se determinar a probabilidade de x sucessos em x ensaios dada qualquer seqüência (ou seja, ao todo), precisamos determinar quantas seqüências específicas são compostas de x "sucessos" em n ensaios. Esse número de seqüências nos é dado pela combinação de n elementos x a x:

$$n\ x = n!x!n\text{-}x!$$

Assim, temos que o número de seqüências com dois filhos homens entre um total de três filhos é

$$3\ 2 = 3!2!3\text{-}2!{=}3{\times}22{=}3$$

Podemos então estabelecer que a probabilidade de se obter x "sucessos em n tentativas é

$$px{=}n\ x\ \ x1\text{-}\pi n\text{-}x$$

Seguindo no exemplo visto acima, a probabilidade de se obter dois filhos homens em um total de 3 filhos é

$$px{=}2{=}3\ 2\ 122123\text{-}2{=}3{\times}1412{=}38$$

O que pode ser verificado consultando-se o espaço amostral correspondente.
Podemos então construir a distribuição de probabilidades de X, correspondente a esse exemplo.

Número de filhos homens: x	Probabilidade px

0	p0=3 0 120123=1×1×18=18
1	p1=3 1 121122=3×1214=38
2	p2=3 2 121122=3×1412=38
3	p3=3 3 123120=1×18×1=18

Pode-se demonstrar que uma variável aleatória com distribuição binomial tem

e

o que pode ser comprovado especificamente para nosso exemplo

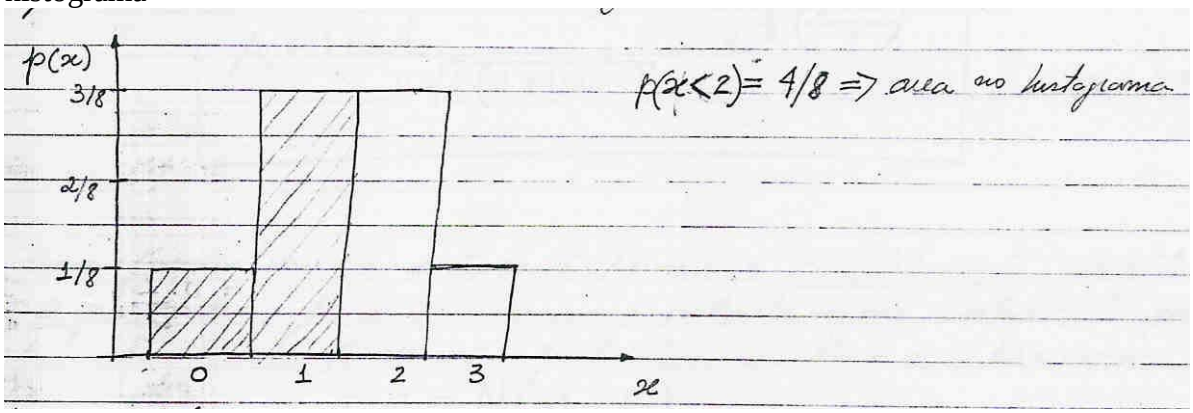
x	px	x px	x-X	x-X2	x-X2 px
0	18	0	-1,5	2,25	2,25×18
1	38	38	-0,5	0,25	0,25×38
2	38	68	+0,5	0,25	0,25×38
3	18	38	+1,5	2,25	2,25×38
∑	1	128	-	-	68

X=x×px=128=1,5⇔X=n=312=1,5

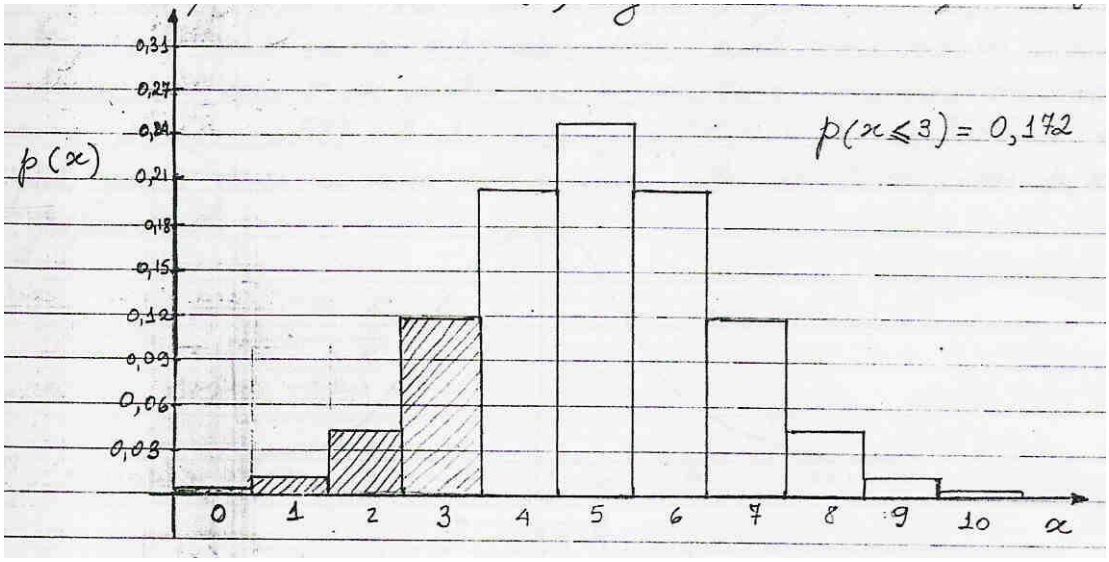
X2=x-X2px=68=34⇔X2=n1-=3×1212=34

II – Variáveis Aleatórias Contínuas: A Distribuição Normal

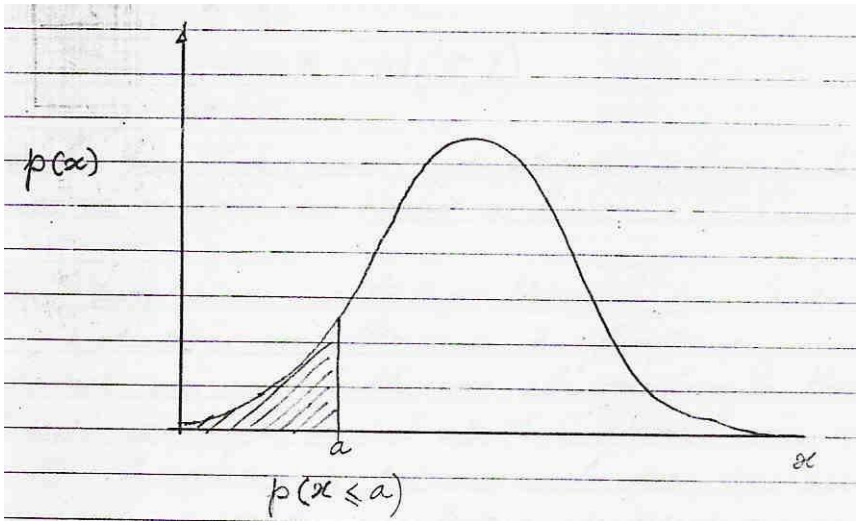
Voltando ao exemplo da variável aleatória "número de filhos homens" em famílias de tamanho (n) igual a três, podemos representar a função de probabilidade através de um histograma



No caso de famílias de tamanho (n) dez esse mesmo histograma seria



Assim, conforme n aumenta, a distribuição de probabilidades se aproxima de uma curva contínua, na qual as probabilidades são apresentadas como áreas debaixo dessa curva:



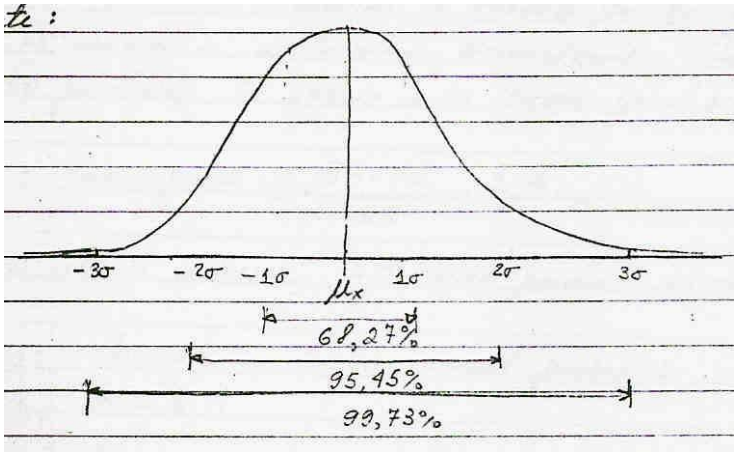
No caso particular em que 12 e que o n é grande, a distribuição binomial se aproxima de uma distribuição de probabilidade contínua de nominada Normal, caracterizada pela seguinte função de probabilidade:

e=2,71828...

Como podemos ver, a distribuição normal é completamente caracterizada por dois parâmetros, X e X, havendo portanto uma distribuição particular para cada combinação de valores particulares de X e X. Escrevemos

X=N(X, X)

A distribuição normal tem algumas propriedades importantes: sua média é X e sua variância é X2 ; além disso, existe uma relação fixa entre e px. Assim, a um de distância à esquerda e à direita da média X temos uma px=.6827. A 2 essa probabilidade é de 0,9545 e a 3σ temos que quase toda a área sob a curva está incluída, com px=.9973 .
Graficamente:



Assim, por exemplo, se tivermos uma variável aleatória X tal que

$$X=N(5,1)$$

Sabemos que temos uma probabilidade inferior a 1% de observarmos valores de x maiores ou iguais a 8.

Como temos uma distribuição normal para cada valor particular de X e/ou X, e dada a complexidade da função de probabilidade normal, é impraticável que se trabalhe com ela diretamente. Para se calcular as probabilidades de uma variável aleatória normal com valores particulares de X e X, transforma-se primeiramente essa variável aleatória para uma variável medida em termos de unidades de desvio-padrão, operação chamada de Padronização. Ou seja, aplica-se a seguinte transformação linear na variável aleatória X:

que como vimos no capítulo de transformações lineares tem

$$Z=0 \quad \text{e} \quad Z=1$$

Uma vez efetuada essa transformação, consulta-se a curva normal que tem essas propriedades (curva essa que já se encontra tabulada), ou seja,

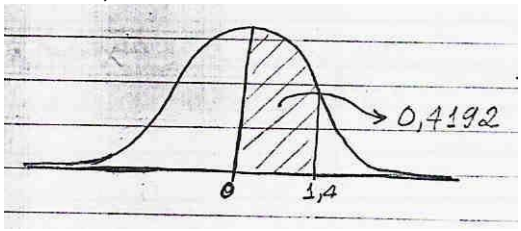
$$Z=N(0,1)$$

e a partir dessa distribuição, chamada Normal Padrão, calcula-se as probabilidades em que se está interessado.

Exemplo: Num certo exame a média foi 72 e o desvio-padrão foi 15, com as variáveis se distribuindo normalmente. Qual a probabilidade de um aluno qualquer ter obtido um exame igual ou superior a 93.

1º passo – Padronização: $Z=93\text{-}7215=1,4$
Esse aluno estaria a 1,4 desvios-padrão da média

2º passo: Consulta à curva normal padrão:
Se $Z=1,4$



$$p0\leq Z\leq 1,4=0,4192$$

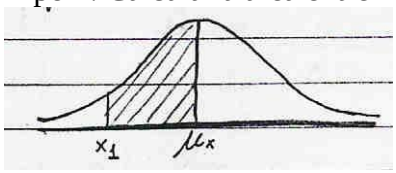
3º passo: Cálculo da probabilidade em questão

$$pZ\geq 1,4=0,5\text{-}0,4192=0,0808 \quad \text{ou} \approx 8,1\%$$

Problemas típicos envolvendo a Distribuição Normal:

Examinemos esses problemas típicos através de um exemplo numérico simples.
Suponhamos: $N(200,10)$

Tipo 1: Calcular a área entre X1 e X



Exemplo: Que proporção de casos tem escores entre 180 e 200?

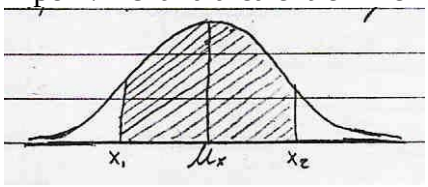
$$Zi=180\text{-}20010=10$$

$$pZ=0,4773$$

R: 0,4773

Solução: área entre X1 e X é dada diretamente pela tabela da curva normal.

Tipo 2: Achar a área entre X1 e X2 (quando estão em lados opostos a X)



Qual é a probabilidade de se obter um escore entre 190 e 212?

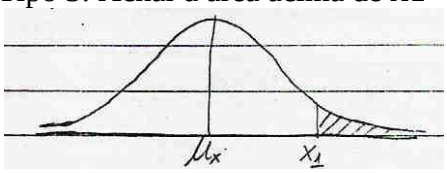
$$Z1=190\text{-}20010=\text{-}1 \quad pZ1=0,3413$$

$$Z2=212\text{-}20010=1{,}2 \quad pZ2=0{,}3849$$

R: 0,3413 + 0,3849 = 0,7262

Solução: Achar a área entre X1 e X e somar a área entre X e X2

Tipo 3: Achar a área acima de X1



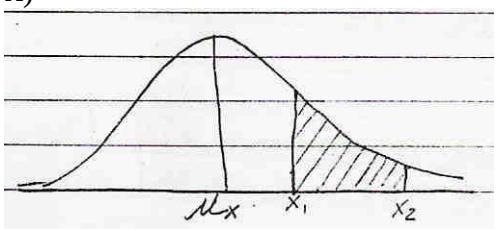
Que percentagem de casos tem escore acima de 227,3?

$$Z1=227{,}3\text{-}20010=2{,}73 \quad pZ1=0{,}4968$$

R: 0,5 - 0,4968 = 0,0032 ou 0,32%

Solução: Achar a área entre X e X1 e subtraí-la de 0,5

Tipo 4: Achar a área entre X1 e X2 (quando X1 e X2 estão do mesmo lado em relação à X)



Qual é a área da curva entre os escores 208 e 222?

$$Z2=222\text{-}20010=2{,}2 \quad pZ2=0{,}4861$$

$$Z1=208\text{-}20010=0{,}8 \quad pZ1=0{,}2881$$

R: 0,4861 - 0,2881 = 0,1980

Solução: Achar a área entre X e X2 e dela subtrair a área entre X e X1

Problema II:
Dados: QI dos estudantes da USP: N (110,10)
QI dos estudantes do IUPERJ: N (130,20)

a. Que proporção dos alunos do IUPERJ tem QI's
   1. Acima de 149?
   2. Entre 129 e 157?
   3. Abaixo um escore Z de +1,75?
   4. Acima do 64° percentil?
   5. Entre 110 e 120?
   6. Com um escore pelo menos 10% acima da média?

b. Um estudante no quinto percentil no IUPERJ estaria em que percentil na USP?

c. Que proporção de estudantes da USP tem QI's acima do QI corresponde ao 33° percentil no IUPERJ?

II.2. Aproximação da Binomial pela Normal

Quando x é grande e nem π nem (1-π) estão muito perto de zero, a distribuição binomial pode ser bastante bem aproximada por uma distribuição normal com a variável padronizada sendo dada por

A aproximação melhora conforme n aumenta, sendo exata no limite. Na prática a aproximação é muito boa se n e n(1-) são maiores do que 5.

Exemplo: (Tirado de M. Spiegel)
A distribuição de probabilidades de "número de caras" em 10 jogadas de uma moeda é, segundo a binomial:

N° de caras	Probabilidades (com =12)
0	1/1024 = 0,001
1	10/1024 = 0,010
2	45/1024 = 0,044
3	120/1024 = 0,117
4	210/1024 = 0,205
5	252/1024 = 0,246
6	210/1024 = 0,205
7	120/1024 = 0,117
8	45/1024 = 0,044
9	10/1024 = 0,010
10	1/1024 = 0,001

A probabilidade de se obter entre 3 e 6 caras é

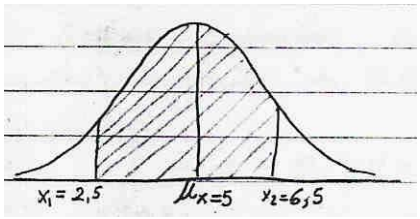
$$P3+P4+P5+P6=0,0773$$

Supondo que os dados são contínuos (isto é, usando a curva normal) segue-se que 3 a 6 caras podem ser consideradas como 2,5 a 6,5 caras.
A média e o desvio-padrão serão

$$X=n=1012=5$$

$$X=n1212=104=1,58$$

Então:



$X1=2,5 \Rightarrow Z1=2,5-51,58=-1,58$

$\therefore pZ1=0,4429$

$X2=6,5 \Rightarrow Z2=6,5-51,58=0,95$

$\therefore pZ2=0,3289$

$\therefore pZ1+pZ2=0,7718$

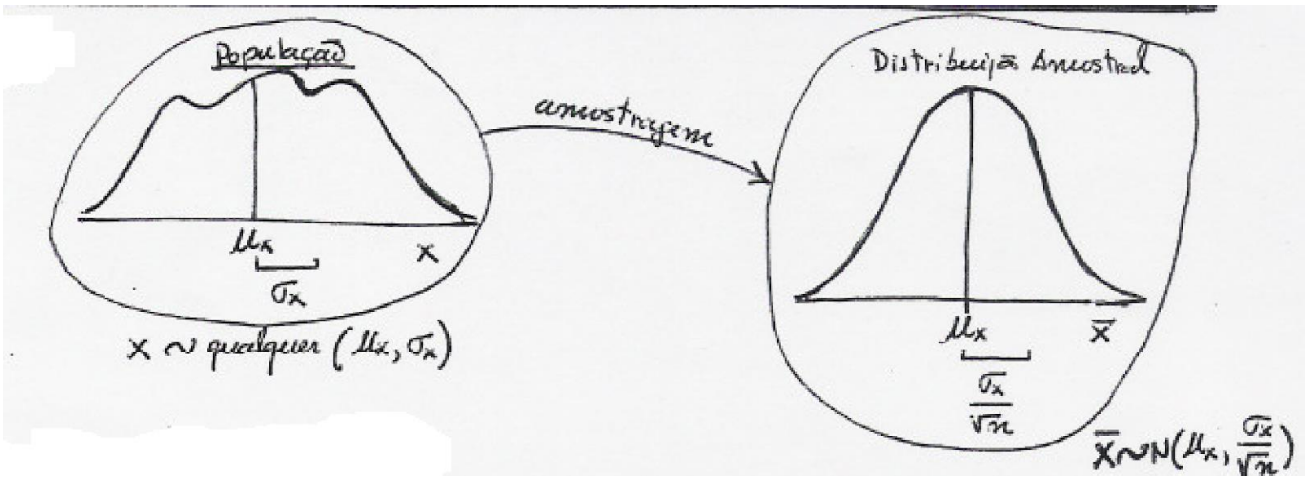
$\therefore p2,5<X<6,5=0,772$
que comparado com o valor exato obtido via binomial (0,773) representa uma aproximação muito boa.

CAPÍTULO 7 – AMOSTRAGEM E INFERÊNCIA

Metodologia – Amostragem, Inferência etc.

1. A Inferência Estatística é baseada no importantíssimo Teorema do Limite Central que, posto em linguagem corrente, pode ser enunciado da seguinte maneira:

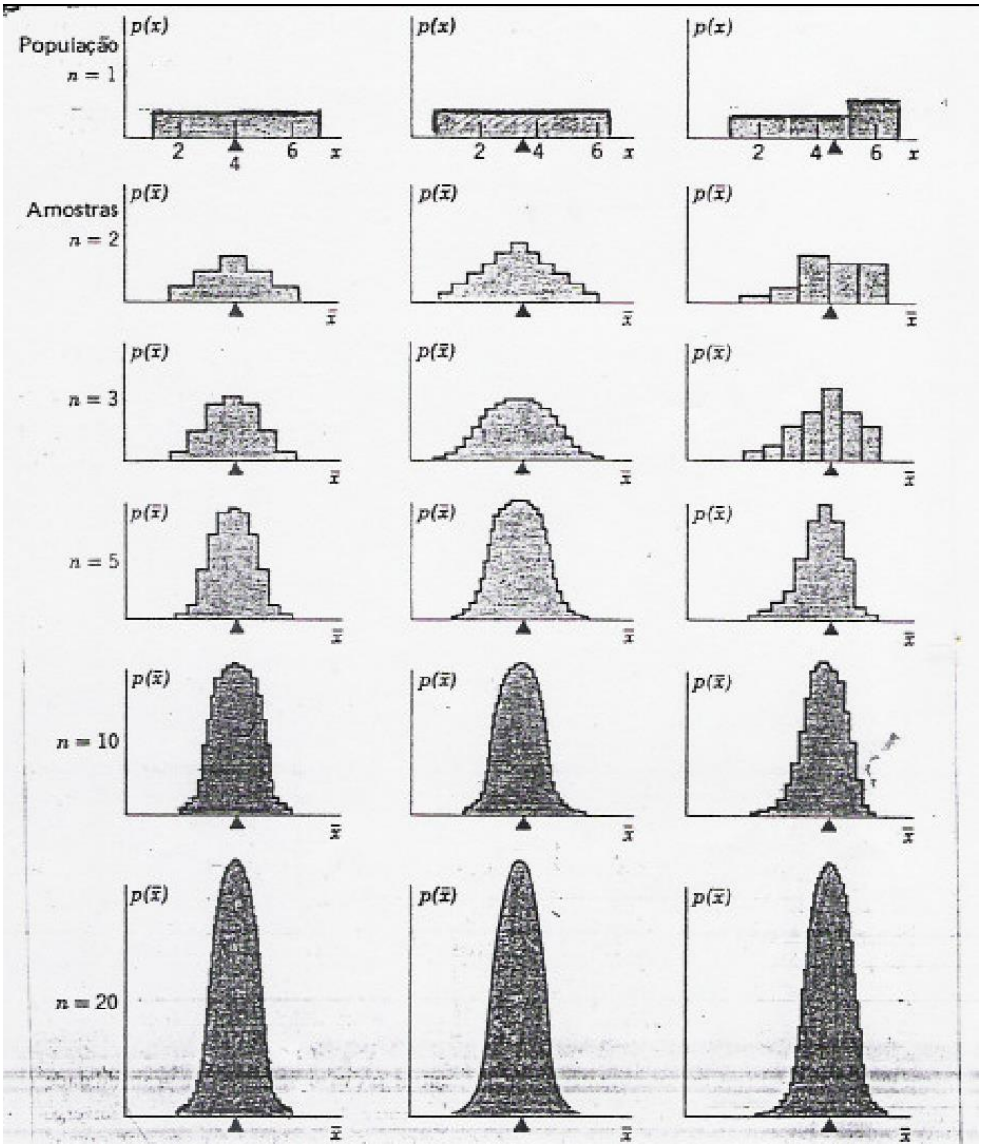
Ou seja,



Observe 2 pontos importantes:

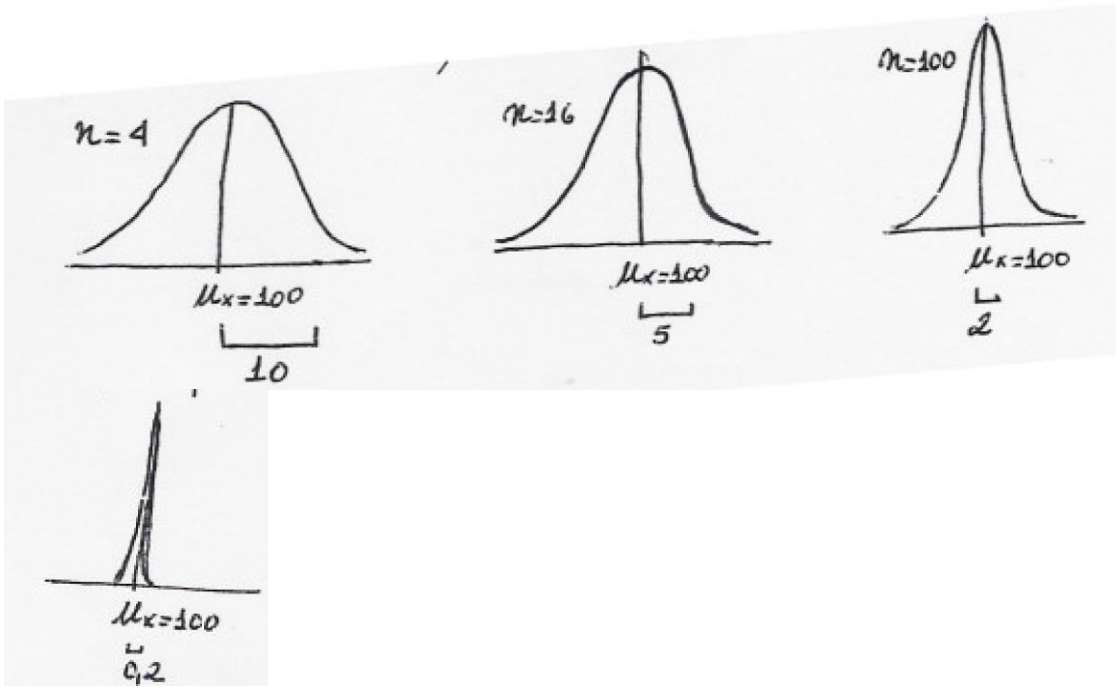
a. A distribuição da Média Amostral tende para a distribuição Normal, quando n aumenta, independentemente do tipo (isto é forma) da distribuição na População

Visualmente:



b. Na Distribuição Amostral a variação em torno da média X, de amostra para amostra (teóricamente), é n vezes menor do que a variação na população, de tal forma que, conforme n cresce, a variação de amostra para amostra diminui rapidamente. No limite, quando a amostra é grande, a média de todas as amostras (i. e., de qualquer amostra) é idêntica a da população.

Por exemplo:
Seja X=100 e X=20
Tirando amostras de tamanho **n** desta população as distribuições amostrais para diversos n's teriam as formas



Agora, quando n=10.000 implica em dizer que 99,7% das amostras terão médias compreendidas entre 99,4 e 100,6. Nessa situação, qualquer amostra (n=10.000) estará com a média praticamente na mosca.
(Lembre-se que a probabilidade numa Normal de X estar distante +- 3 desvios da média é de 99,7%)

2. Problema: Se na população X tem média 500 e desvio-padrão 100, qual a probabilidade de, tendo-se tirado uma amostra de 25 individuos, a média de minha amostra estar entre 480 e 520?

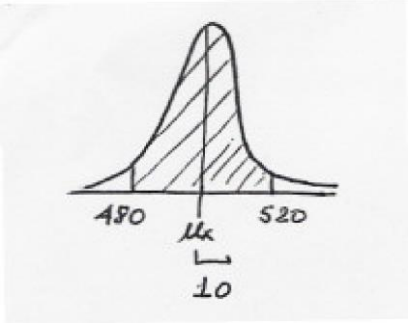
População:

$$Z1=480-50020=-1\Rightarrow pZ1=0,3413$$

$$Z2=520-50020=+1\Rightarrow pZ2=0,3413$$

Resposta: Ou 68,26% das amostras possíveis.

Suponhamos agora que n=100. Qual então seria a resposta?



$$Xn=100100=10$$

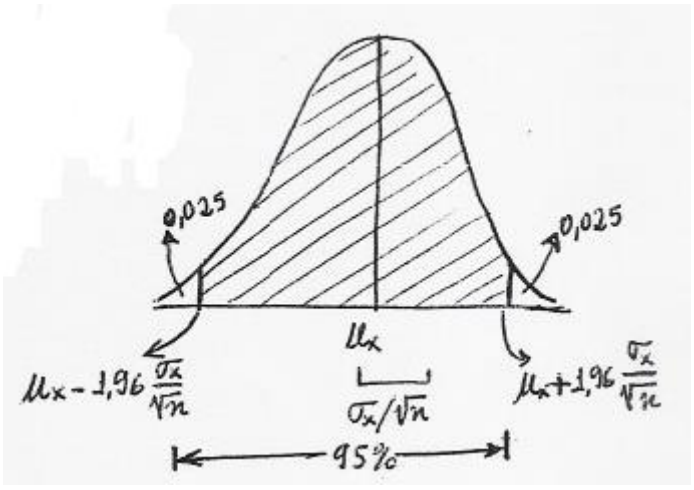
$$\therefore Z1=480\text{-}50010=\text{-}2\Rightarrow pZ1=0,4772$$

$$Z2=520\text{-}50010=+2\Rightarrow pZ2=0,4772$$

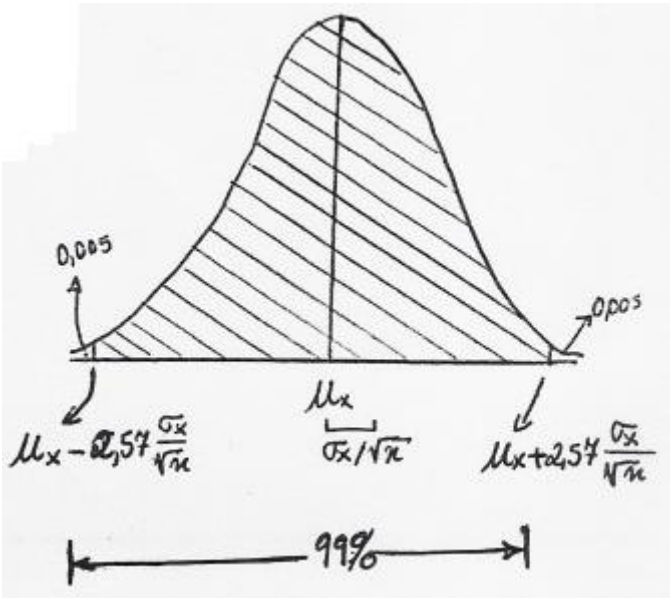
Resposta:

Ou 95,44% das amostras possíveis de tamanho 100 terão médias entre 480 e 520.

3. Pelo que então sabemos a respeito da Normal, podemos escrever



E



Grave bem estas relações!

Nota: Por razões óbvias, Xn denomina-se "êrro-padrão de estimativa".

Suponhamos que eu esteja observando um conjunto de 5 indivíduos. A cada um pergunto "em quem vai votar". As respostas a essa pergunta eu codifico da seguinte forma:

=1 se for votar no Brizola e

=0 se for votar em outro candidato

Suponhamos agora que obtive as seguintes respostas: eleitores: X=1,0,1,1,0.

a. Calcule a <u>média</u> de X;
b. Qual a proporção votando em Brizola?

Resposta: X=média=1+0+1+1+05=35=0,6

p=proporção votando em Brizola=35=0,6

Ou seja: e vice-versa.

O desvio-padrão dessa variável é também fácil de calculas, sendo dado por

Onde p é a proporção. No nosso exemplo p=0,6, logo

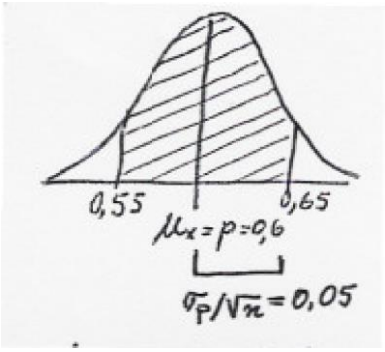
$$p=0,6\ 1-0,6=0,60,4=0,24=0,49$$

Podemos, pois proceder normalmente como no caso de uma média qualquer.

Por exemplo, podemos perguntar, "Qual a probabilidade de, tirando uma amostra de tamanho n=100, obter uma proporção amostral de eleitoras do Brizola entre 0,55 e 0,65, sendo a proporção na população igual a 0,6?

Resposta: Distribuição Amostral da Proporção:
Sendo:

$$pn=0,49100=0,4910\approx0,05$$



Como 0,55 e 0,65 definem o intervalo +- 1 desvio-padrão em torno da média, então sabemos que essa probabilidade é de 0,6826 (conforme vimos no problema anterior).

Se n fosse 10.000 ao invés de 100, o êrro padrão da proporção amostral seria apenas de

$$pn=0,4910000=0,49100=0,0049\approx0,005$$

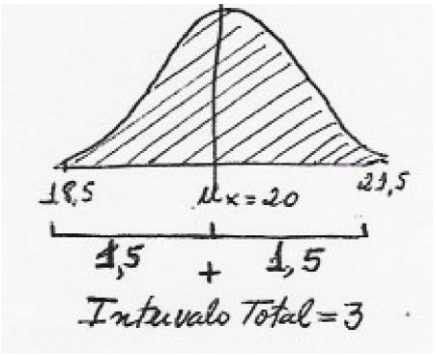
Considerando que entre +3 e -3 desvios (êrros)-padrão da média temos praticamente todos os casos incluídos (na verdade 99,7%), podemos dizer que "sendo a proporção de votantes de Brizola na população igual a 0,6, se tirarmos uma amostra de 10.000 eleitores, temos praticamente certeza de que a proiporção nessa amostra vai estar entre 0,585 e 0,615 (isto é, entre 0,6-3(0,005) e 0,6+3(0,005))". Ou seja, tenho um êrro total de apenas 1,5% para cima ou para baixo, no máximo. Provavelmente, o êrro que estarei cometendo com uma amostra desse tamanho é bem menor que isso.

Como você pode ver, estou quase fazendo inferências a partir da amostra...

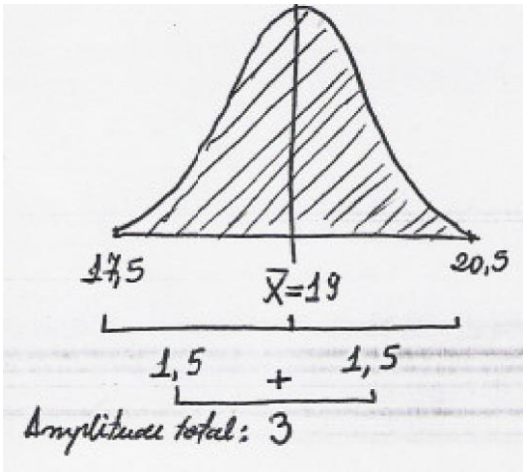
5. Imaginemos agora a seguinte situação: População: $X=20$ e $X=5$

Então eu tenho praticamente certeza de que, com uma amostra de n=100, a minha média amostral vai estar entre $20-35100$ e $20+35100$, ou seja, entre 18,5 e 21,5.

Distribuição de X:

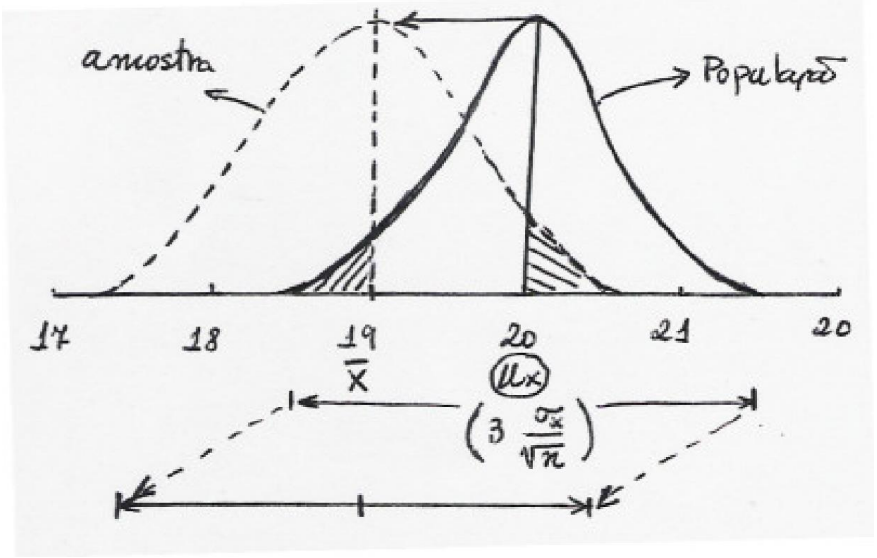


Suponhamos agora que eu tenha tirado uma amostra e que tenha verificado que X=19. Então, se eu sei que todas as amostras estavam incluídas em um intervalo de 1,5 em torno da média X, posso ter certeza que essa média da população estará incluída num intervalo de 1,5 em torno da média da amostra.



Ou seja, estará incluída no intervalo 17,5 a 20,5.
Visualmente, isso equivale a deslocar o intervalo, centrando-o na média amostral

---



Nota: Observe que a "probabilidade de se obter X=19 ou mais sendo a média X" é igual à "probabilidade de X=20 tendo-se obtido X=19" (dadas pelas áreas observadas acima).

Isso nos permite então inverter as relações vistas anteriormente, escrevendo

Finalmente! Agora podemos dizer algo a respeito da média da amostra (conhecida): ou seja, somos capazes fazer Inferência.

6. Observe que

e "Erro de estimativa (total)"

definem os chamados "Intervalos de Confiança" de 95% e 99% de probabilidade, respectivamente, para a estimativa da média da população.

Exemplo: Suponhamos que para uma certa população saibamos que seu desvio-padrão é X=3. Extraindo uma amostra de n = 100 indivíduos obtemos uma média amostral de 18. Qual o intervalo de confiança a 99% para a estimativa da média da população?
Resposta: O intervalo de confiança a 99% é dado por

$$X=X\pm \text{êrro de estimativa}=X2{,}57Xn$$

$$=18\pm 2{,}573100=18\pm 0{,}77\approx 18\pm 0{,}8$$

Ou seja, temos 99% de certeza que a média da população está entre 17,2 e 18,8, aproximadamente. O nosso êrro de estimativa é de $\pm 0{,}77$ cm 99% de confiança.

7. Como você já deve ter notado, a nossa regra de inferência depende de conhecermos X, o desvio-padrão na população, que obviamente é também desconhecido. Como sair dessa enrascada? Acontece que se pode provar que o desvio-padrão amostral SX (com aquela correção de dividir por N-1 ao invés de por N, como já vimos) é a melhor

estimativa que podemos fazer de X. Assim, trata-se simplesmente de substituir nas expressões anteriores X por SX!

Exemplos:

a. Uma amostra de 100 homens tirada de uma certa população indicou que a altura média (em polegadas) dessa amostra é 71 cm com variância 9.

Resposta: O intervalo de confiança a 95% para a estimativa da média é

$$X=X\pm\text{êrro de estimativa}=X1,96Xn$$

Estimando X por SX, e substituindo na expressão acima, obtemos

$$SX=SX2=9=3$$

$$X=X1,,96SXn=71\pm3100*1,96=71\pm0,6$$

Ou seja, temos 95% de confiança que a média da população está entre 70,4 e 71,6.

b. Pouco antes da eleição presidencial de 1972 nos EUA, uma pesquisa Gallup feita junto a 2000 eleitores acusou 760 favoráveis ao McGovern e 1240 favoráveis a Nixon. Calcular o intervalo de confiança de 95% para a proporção na população (π) que votou a favor de McGovern.

Resposta: n=2000 e a proporção na amostra é p=760/2000=0,38

O desvio-padrão, como já sabemos, no caso de uma proporção é função de p:

$$Sp=p1-p=0,381-0,38=0,236$$

Portanto, a estimativa da proporção na população π num intervalo de confiança de 95% é dada por

$$\pi=p\pm1,96Spn=0,38\pm1,960,2362000\approx0,38\pm0,02$$

Ou seja, temos 95% de certeza de que a proporção de eleitores que votarão em McGovern está entre 36 e 40% da população. O êrro de estimativa total é de $\pm2\%$.
(De fato, a proporção dos que votaram em McGovern naquela eleição foi de 38,2%)

Se quizéssemos fazer uma estimativa com 99% de confiança, teríamos que fazer

$$\pi=p\pm2,57Spn=0,38\pm2,570,2362000\approx0,38\pm0,03$$

Ou seja, temos 99% de certeza de que a proporção de eleitores que votarão em McGovern está entre 35 e 41% da população. O êrro de estimativa é agora de $\pm3\%$.
É assim que as estimativas das prévias são, ou melhor, deveriam ser feitas.

8. Finalmente, algo também muito importante:

Observe que a expressão para o erro de estimativa a um nível de confiança de 95% é

$$\text{êrro}=1,96Xn$$

(Valor de Z correspondente a 95%: Z95%)

Para 99% de confiança é

$$\hat{e}rro = 2{,}57 X n$$

(Valor de Z correspondente a 99%: Z99%)

Ou seja, de modo genérico escrvemos:

Isso nos permite escrever que n, o tamanho da amostra é dado por

$$n = Z\% X \hat{e}rro \Longrightarrow$$

Observe que o tamanho da amostra pode ser visto como função de

1. O nível de confiança que eu quero ter na minha inferência: quanto mair o nível de confiança (isto é Z%), maior o tamanho da amostra necessária;
2. A heterogeneidade da população (isto é X):quanto mais heterogâneo o fenômeno que eu estiver querendo observar, maior a amostra necessária para observá-lo;
3. O êrro (absoluto) que eu estou disposto a incorrer: quanto maior o êrro que eu quizer admitir, menor a amostra que eu preciso.

Assim, o tamanho da amostra n não tem nada a ver com o tamanho da população. Ele é função apenas do nível de confiança, da heterogeneidade da população e do êrro de estimativa que eu quero cometer. Portanto, a rigor, não faz muito sentido (teórico) dizer "vou tirar uma amostra de 5% da população" ou qualquer outra coisa parecida.

9. Em termos práticos, a expressão $n = Z\% X \hat{e}rro^2$ nos dá a maneira de calcular o tamanho da amostra quando vou fazer um levantamento.

Por exemplo, suponhamos que eu saiba (através do Censo, por exemplo) que a média de escolaridade na minha população é de 8 anos e o desvio é de 4 anos. Eu quero fazer uma pesquisa e só admito um erro máximo na minha estimativa de 0,5 ano. Pretendo trabalhar com um nível de confiança de 95%. Qual o tamanho da amostra que eu necessito para a minha pesquisa?

Resposta: $n = 1{,}96 * 40{,}52 \approx 246$
Ou seja, necessito de uma amostra de aproximadamente 246 indivíduos. A um nível de confiança de 99% seriam necessários

$$n = 2{,}57 * 40{,}52 \approx 423 \text{ indivíduos}$$

10. Imagine agora o caso de uma prévia eleitoral. Qual o número de casos na minha amostra necessários para fazer inferência a 95% de confiança e incorrendo num erro máximo de 3%?

Resposta: observe que a pior situação possível é aquela em que a proporção votando num candidato é 50%: nesse caso, o desvio-padrão (a heterogeneidade) é máximo e, portanto, configura uma situação limite para o tamanho da amostra.

Vamos, portanto, supor que, na pior das hipóteses, p=0,5. Logo:

$$Sp=0,51-0,5=0,25=0,5$$

Assim,

$$n=1,96*0,50,032\approx1067 \text{ indivíduos no máximo.}$$

Se eu estou disposto a admitir um êrro maior, digamos $\pm 4\%$, o valor de n será

$$n=1,96*0,50,042\approx600 \text{ indivíduos no máximo.}$$

CAPÍTULO 8 – TESTE DE HIPÓTESE (E ESTIMAÇÃO COM VARIÂNCIA DESCONHECIDA) – (3 AULAS: TH1, TH2 E T-STUDENT)

INCLUIR AULA TESTE DE HIPÓTESE 1 (ESTÁ EM PDF)

Curso Lego I: Introdução à Análise de Dados
Profº Nelson do Valle Silva
Aula 8

Testes de Hipóteses (II)

Nos testes de hipóteses o problema básico consiste em se testar a plausibilidade de uma (ou mais) hipóteses a respeito do parâmetro populacional, tendo em vista os resultados obtidos numa amostra particular.
Um exemplo talvez nos permita apresentar os conceitos e operações principais envolvidas nos testes de hipóteses:
Suponhamos que você costume disputar o "chopinho" com um amigo na base de um jogo de "cara ou coroa". Dando "cara" você perde, pagando a conta. Após 100 partidas você observa que pagou o "chopinho" aproximadamente 60% das vezes.Estranhando esta proporção que lhe é altamente desfavorável, você se confronta com duas opções alternativas:

1. Você tem um grande azar no jogo de "cara ou coroa", esta proporção refletindo o seu azar;
2. É possível que seu amigo esteja de fato o enganando, e a moeda que vocês têm usado (sempre fornecida por ele) é, de fato, uma moeda "viciada", daquelas que você sabe que estão à venda e que tem uma probabilidade de sair "cara" igual a ¾.

Obviamente, você tem que decidir entre as duas opções, sendo que a rejeição de uma implica na aceitação da outra, decisão essa que vai afetar ou não sua amizade por seu "amigo". Observe que a primeira opção é a de que seu amigo é honesto, ou seja, que a moeda é equilibrada com probabilidade de "cara" igual a de "coroa" (isto é ambas igual a ½). Essa hipótese de normalidade no resultado pode ser chamada de hipótese nula e podemos escrever

$$H0: \pi=12=0,50$$

A outra hipótese, a da desonestidade da moeda, pode ser chamada de hipótese alternativa, sendo escrita

$$H1: \pi=34=0,75$$

Supondo-se que a hipótese nula é verdadeira, você tem então que

$$p= \pi=12=0,50$$

E

$$p=\pi(1-\pi)100=0,5(0,5)100=0,050$$

Isso implica que a probabilidade de você pagar 0,60 ou mais dos "chopps" é dada por

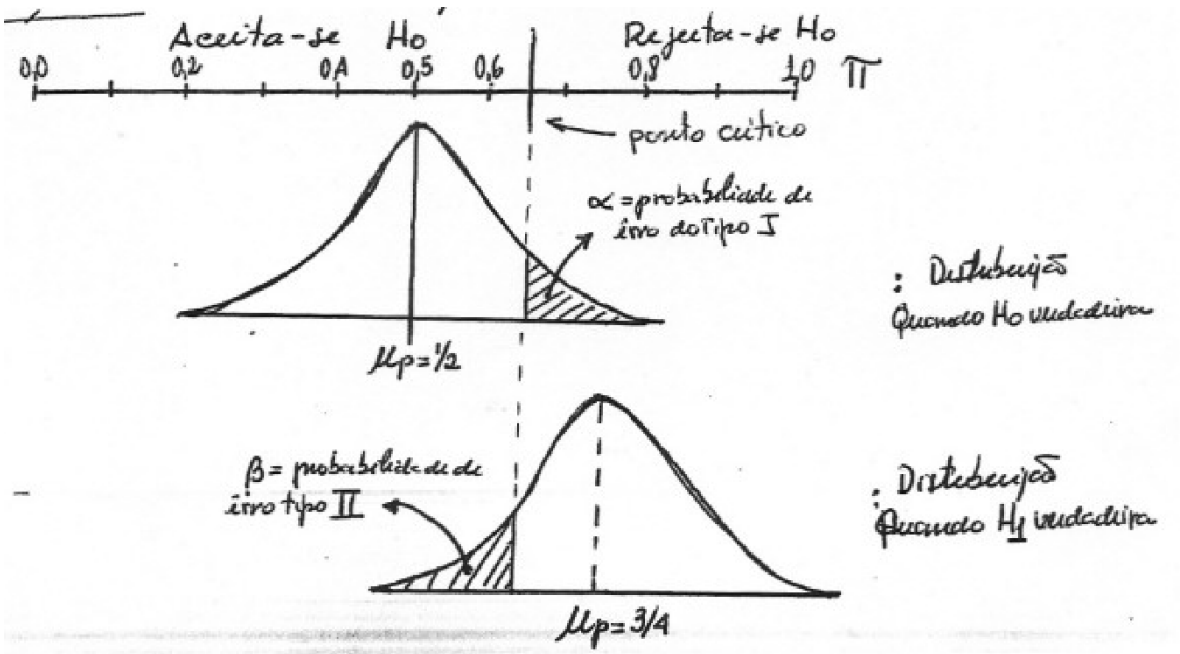
$$Zp=0,60-0,500,050=0,100,05=2,0$$

Claramente você vai ter que estabelecer uma regra de decisão a respeito da menor proporção admissível, valores inferiores a este limite o levarão a rejeitar a idéia do "azar", e naturalmente, aceitar a hipótese da desonestidade. Suponhamos que você, intuitivamente, estabeleça este limite em p=0,65. Logo

$$p \geq 0{,}065 = 0{,}0013$$

Chamemos o ponto limite estabelecido de ponto crítico, e a probabilidade de se obter um valor acima deste ponto crítico sendo H_0 verdadeira de α. Valores com probabilidades inferiores a α, que levam naturalmente a rejeitar H_0, são chamados de estatisticamente significantes. Observe que α corresponde à probabilidade de se rejeitar H_0 quando ela é verdadeira, ou seja, de se cometer um erro chamado erro de tipo I.
Mas, e se H_0 não for verdadeira? Observe que a imposição de um ponto crítico para a rejeição implica não só no estabelecimento de uma probabilidade para o erro de tipo I mensurado acima (ou seja, α, a probabilidade de rejeitar H_0 sendo ela verdadeira), como também no estabelecimento de um erro de aceitar H_0 sendo ela falsa. Chamemos essa probabilidade (de erro de tipo II) de β.

Graficamente:



Quando estabelecemos que o ponto crítico é 0,65 (α=0,0013) temos que a probabilidade de erro do tipo II, ou seja, aceitar H_0 quando ela é falsa e, portanto H_1 verdadeira, nos é dada por

$$p = \pi = 0{,}75$$

E

$$p = 0{,}75(0{,}25)100 = 0{,}043$$

Logo:

$$Zp = 0{,}65 - 0{,}750{,}043 = 0{,}100{,}043 = -2{,}33$$

E

$$p \geq 0{,}65 \approx 0{,}0099 \approx 0{,}01$$

B é, portanto, a probabilidade de se cometer o erro de tipo II, aceitar H_0 quando H_1 é verdadeira (e, assim, H_0 é falsa). Podemos então condensar os conceitos envolvidos nos testes de hipóteses no seguinte quadro:

Realidade	**Aceitar H_0**	**Rejeitar H_0**
H_0 verdadeira; H_1 falsa	Decisão Correta. Probabilidade = 1-α é o "nível de confiança"	Decisão Errada. Probabilidade = α é o "nível de significância"
H_0 falsa; H_1 verdadeira	Decisão Errada. Probabilidade = β	Decisão Correta. Probabilidade = 1-β Chamada de "potência do teste"

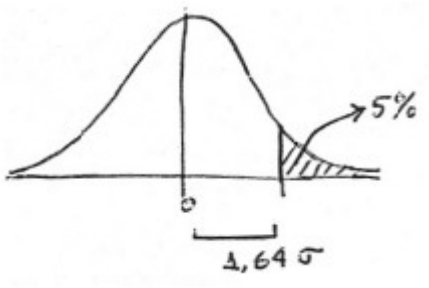
É claro que a rejeição de H_0 traz um custo (a perda da amizade, por exemplo) e como tal α deve ser minimizado. Por outro lado, queremos minimizar também β, uma vez que a continuação do jogo implica em custos monetários. No entanto, esses são objetivos contraditórios, uma vez que quanto menor α maior β (e vice-versa, como pode ser visto na figura). A única maneira de se reduzir simultaneamente α e β é aumentando o tamanho da amostra n, o que segue diretamente das equações de erro padrão. Fora dessa possibilidade, a decisão vai depender essencialmente do custo relativo de se cometer cada tipo de erro.

De uma forma geral os erros de tipo I são mais graves e, assim, se estabelece costumeiramente valores pequenos para α – sendo estes convencionalmente fixados em 5% ou 1%. Assim, o teste de hipóteses envolve normalmente 3 etapas:

1. Estabelece-se a hipótese nula H_0, a hipótese alternativa H_1, bem como o tamanho da amostra e o nível de significância;
2. Assume-se que a hipótese nula é verdadeira. Calcula-se o valor (ponto) crítico correspondente ao nível de significância estabelecido. Por exemplo, se tivermos estabelecido α=5%, sabemos que numa curva Normal se

$$pZ=0{,}05 \Longrightarrow Z \approx 1{,}64$$

Então o valor crítico para p no exemplo acima é



$$Z\text{crítico}=P-\pi\pi(1-\pi)n=1{,}64$$

Para H_0: $\pi=1/2$ temos:

$$p-0{,}50{,}05=1{,}64 \Rightarrow p\text{crítico}=1{,}64*0{,}05+0{,}5=0{,}582.\text{Ou } 58{,}2\%$$

Dessa forma proporções acima de p=0,582 levarão à rejeição da Hipótese nula (com probabilidade = α=5% ou menos).

3. Dados os valores da amostra, uma decisão é tomada. Se o p observado for maior que o valor crítico então H_0 é rejeitada e H_1 é aceita. Caso contrário, não rejeitamos H_0. No nosso exemplo, a proporção observada foi p=0,60. Logo, comparando com o ponto crítico $p_{crítico}=0{,}582$, rejeitamos H_0 e aceitamos H_1, ou seja, rejeitamos a hipótese que temos "azar" no jogo da moeda e que, portanto, aceitamos a hipótese que o amigo é desonesto, com uma probabilidade de 5% de estarmos equivocados.
4. Alternativamente, como o $Z_{crítico}$ a 95% é 1,64, toda vez que o valor de Z na nossa amostra for maior que 1,64 → rejeita-se a hipótese nula ao nível de significância de 5%. Assim, como no nosso exemplo

$$Z=\frac{0{,}6-0{,}5}{0{,}05}=+2>1{,}64\Rightarrow \text{Rejeita-se H0 a favor de H1}$$

Esta é a maneira mais simples de abordar o teste de hipóteses, isto é, comparando o valor de Z obtido com os valores de Z críticos.

Hipóteses Compostas

No exemplo anterior, especificamos uma hipótese nula e uma hipótese alternativa. Freqüentemente é o caso, porém, de termos um conjunto de hipóteses alternativas e não apenas uma. Por exemplo, poderíamos hipotetizar que a moeda é simplesmente "desonesta", sendo $\pi>0{,}50$. Assim, teríamos

$$H0: \pi=0{,}5$$

$$H1: \pi>0{,}5$$

Sendo H_1 composta de várias sub-hipóteses

$$H1: \pi=0{,}51$$

$$H1: \pi=0{,}52 \text{ etc.}$$

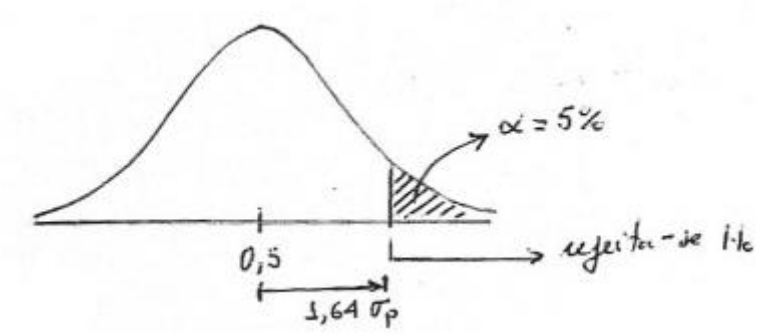
Observe-se que enquanto temos ainda um só valor para α, β nesse caso assume um conjunto de valores possíveis, sendo de fato uma função de π. Essa função é chamada de "função potência".
No entanto, como a avaliação de α permanece sendo simples, o fato de H_1 ser complexo é mais uma razão para nos concentrarmos em avaliar α, segundo o procedimento delineado anteriormente.
Quanto à interpretação, a introdução de hipóteses compostas introduz também uma modificação: como não temos uma hipótese específica H_1 para aceitarmos em substituição a H_0, tudo o que podemos fazer é "não rejeitar H_0" ou "rejeitar H_0", sendo que esta última decisão não implica na aceitação de qualquer hipótese alternativa específica. Especialmente porque como H_1 inclui valores de π extremamente próximos do valor estabelecido em H_0, o nosso risco de incorrermos em erro de tipo II (β) pode ser extremamente elevado.
A solução para esse problema é relatar o valor da probabilidade do valor amostral ter ocorrido na hipótese de H_0 ser verdadeiro, suspendendo-se neste caso o teste da hipótese nula. Assim, no nosso exemplo, ao invés de realizarmos o teste de hipóteses de que H_0: $\pi=0{,}5$, simplesmente relataríamos que a probabilidade de termos uma proporção de perdas igual ou maior que 0,60 sendo que a moeda é honesta (ou seja, H_0 é verdadeira) é igual a 0,0228 ou 2,28%.

Hipóteses de duas caudas

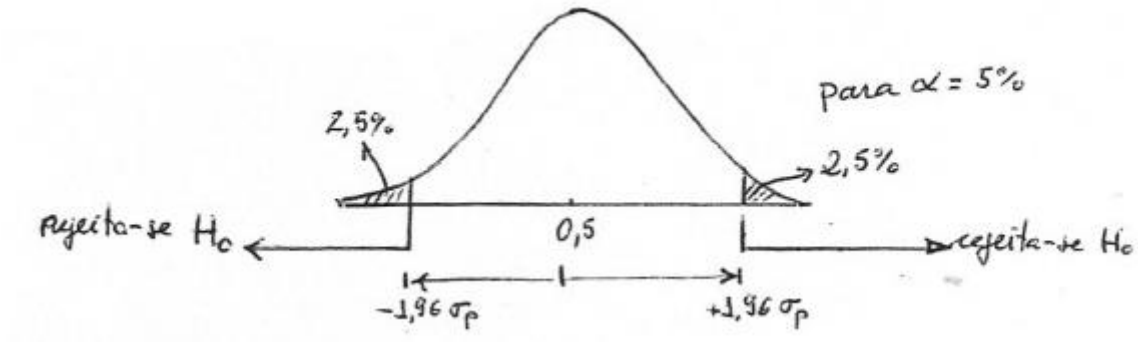
Freqüentemente é o caso em que estamos interessados numa hipótese alternativa sem direção definida. Por exemplo, poderíamos testar para o nosso exemplo a hipótese de que H0: $\pi=0,5$

H1: $\pi\neq0,5$ ou seja H1: $\pi>0,5$ ou H1: $\pi<0,5$

Que corresponderia a hipótese alternativa de que simplesmente a moeda é viciada, podendo tanto o ser no sentido de favorecer a saída de "caras" quanto no sentido de favorecer "coroas". Essa é uma hipótese de duas caudas. Comparando com a situação anterior (1 cauda):



Na nova situação (2 caudas):



Assim, rejeita-se H_0 a α=5% se

$$Z=p-\pi\pi(1-\pi)n>1,96$$

Ou seja, se p cair dentro do intervalo de confiança de 95% H_0 não será rejeitado; se no entanto cair fora do intervalo, rejeitamos H_0.

Para o nosso exemplo em que n=100

$$Z=0,60-0,500,05=2>1,96$$

Ou seja, rejeita-se H_0 em favor de H1: $\pi \neq 0,5$ ao nível de significância de α=5%.

Exercícios Resolvidos a partir de Dados do Human Development Report (1990)

1. Taxa de Crescimento do PNB (1965-1980) entre Países Socialistas e Não-socialistas

Pelo SPSS obtive:

G1 Não-socialistas: X1=2,4558 S1=2,283 n1=113

G2 Socialistas: X1=1,1353 S1=1,511 n1=17

Ao nível de confiança de 95%

$$1\text{-}2 = X1\text{-}X2 \pm 1,9612n1+22n2$$

$$=(2,4558\text{-}1,1353) \pm 1,962,2832113+1,1511217$$

$$=1,3205 \pm 1,960,0461+0,1343$$

$$=1,3205 \pm 1,960,4248=1,3205 \pm 0,08325$$

$$\Pr(0,488 \leq 1\text{-}2 \leq 2,153)=0,95$$

Como o valor zero não faz parte do intervalo → rejeita-se a hipótese nula de não-diferença.

No teste clássico: H0: 1-2=0 ou H0: 1=2

Acredito que a hipótese alternativa deve ser H1: 1-20 ou H0: 1>2

O que é uma hipótese (de um cauda) composta

$$Z=X1\text{-}X212n1+22n2=1,32050,4248=+3,109$$

Comparando com o valor crítico para Z (α=0,95)=1,64, verificamos que a hipótese nula deve ser rejeitada. Como o valor de Z obtido está acima do valor 3, sabemos que a probabilidade de se obter este valor sendo H_0 verdadeira é menor que 0,001.

2. Taxa Bruta de Mortalidade (1988) entre Países Ilhas e Não-ilhas

G1 Não-ilhas: X1=11,6783 S1=4,755 n1=115

G2 Ilhas: X1=8,7333 S1=2,915 n1=15

$$1\text{-}2=(11{,}6783\text{-}8{,}7333)\pm 1{,}964{,}7552115+2{,}915215$$

$$=2{,}945\pm 1{,}960{,}1966+0{,}5665$$

$$=2{,}945\pm 1{,}960{,}8736=2{,}945\pm 1{,}712$$

$$\Pr\,(1{,}233\leq 1\text{-}2\leq 4{,}657)=0{,}95$$

Como zero não faz parte do intervalo → rejeita-se a hipótese nula de não-diferença.

No teste clássico acredito que novamente: H0: $1\text{-}2=0$ ou H0: $1=2$

H1: $1\text{-}2\geq 0$ ou H0: $1>2$

Hipótese (de um cauda) composta

$$Z=X1\text{-}X212n1+22n2=2{,}9451{,}712=+1{,}720$$

Que comparado com o valor crítico para Z_c=1,64 implica na rejeição da hipótese nula. Mais especificamente, a probabilidade de se obter a diferença entre médias acima dada H_0 verdadeira é de 0,0427 ou 4,27%. O que aconteceria se, ao invés de utilizar α=5%, tivéssemos optado por α=1%?

Curso Lego I: Introdução à Análise de Dados
Profº Nelson do Valle Silva
Aula 9

Estimação com Variância Desconhecida

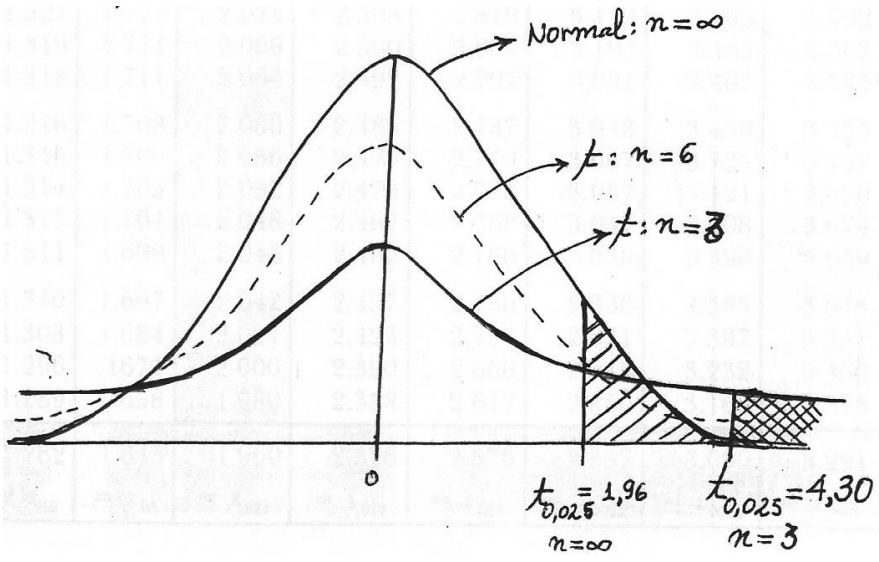
Como já vimos, quando o tamanho da amostra é suficientemente grande, a distribuição amostral de X (isto é, a distribuição em todas as amostras possíveis de um tamanho n dado) é aproximadamente Normal com média X variância X2n. Assim os testes de

hipóteses a respeito de X envolviam o cálculo da estatística:

Essa expressão é absolutamente correta quando X é conhecido. Mas o caso comum é que, de fato, desconhecemos X, tendo este que ser estimado pela estatística da própria

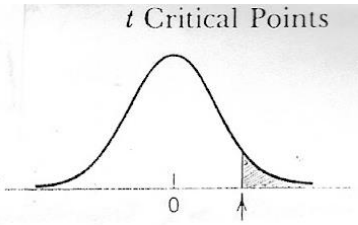
amostra SX. Neste caso, a estatística

se distribui como uma distribuição chamada "t de Student". Esta é a distribuição teoricamente correta para amostras de qualquer tamanho quando X é desconhecido. Apenas a distribuição "t de Student" depende do tamanho da amostra *n*, sendo que quando *n* é grande (isto é, algo maior que 120) a distribuição "t" é praticamente idêntica à Normal. (Daí a aproximação que vínhamos usando até agora).
Graficamente:



A distribuição de “t” já se encontra tabulada. Apenas, ao invés de a termos tabulada segundo os valores de *n*, a tabulação é feita segundo o divisor da variância da amostra SX2 (ou seja, *n-1*). Esse valor é chamado de “graus de liberdade”

t Critical Points



Critical point. For example: $t_{.025}$ leaves .025 probability in the tail.

n-1 ↓

1 cauda | 2 caudas

	5% ↓	5% ↓	1% ↓	1% ↓
d.f.	$t_{.25}$	$t_{.10}$	$t_{.05}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.0025}$	$t_{.0010}$	$t_{.0005}$
1	1.000	3.078	6.314	12.706	31.821	63.637	127.32	318.31	636.62
2	.816	1.886	2.920	4.303	6.965	9.925	14.089	22.326	31.598
3	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.711	1.415	1.895	2.365	2.998	3.499	4.020	4.785	5.408
8	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.257	3.189
22	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.679	1.296	.1671	2.000	2.390	2.660	2.915	3.232	3.460
120	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291
	$= z_{.25}$	$= z_{.10}$	$= z_{.05}$	$= z_{.025}$	$= z_{.010}$	$= z_{.005}$	$= z_{.0025}$	$= z_{.0010}$	$= z_{.0005}$

← Normal

Assim, por exemplo, para uma amostra de tamanho n=3, podemos usar a distribuição "t de Student" para estimar o intervalo de confiança a 95%:

t0,025=4,303 para n-1=3-1=2 graus de liberdade

Logo,

$$\mu = X \pm 4{,}303 S_{Xn}$$

Observe-se que toda vez que ignoramos X a distribuição "t de Student" é a distribuição adequada para estabelecer os intervalos de confiança e fazer os testes de hipótese. No entanto, como já vimos, para amostras grandes, a curva Normal nos dá uma aproximação bastante satisfatória, uma vez que a distribuição "t" converge para uma Normal Padrão conforme $n \to \infty$.

Exemplo: Uma amostra de 9 mensurações do diâmetro de uma esfera indicou uma média de $X = 4{,}38$ cm. E um desvio-padrão de $S_X = 0{,}06$ cm. Achar o intervalo de confiança a 99% para o parâmetro verdadeiro.

R: $X = X t_{0{,}005} S_{Xn}$
$t_{0{,}005} = 3{,}355$ para $v = n\text{-}1 = 9\text{-}1 = 8$ graus de liberdade
Logo:

$$X = 4{,}38 \pm 3{,}355 0{,}069 = 4{,}38 \pm 0{,}067$$

O intervalo de confiança é portanto $\Pr 4{,}313 \leq X \leq 4{,}447 = 0{,}99$

(Comparar com o resultado obtido se usássemos a Normal.)

A distribuição "t" é também adequada quando queremos estimar diferenças entre médias. Para isso vai-se supor que ambas as populações tem a mesma variância $X2$. Quando desconhecemos $X2$, temos que estimá-la através das variâncias das amostras $S12$ e $S22$, calculando um valor médio a partir destas, cada uma ponderada por seu respectivo tamanho de amostra:

Para $H0{:}1 = 2 \Longrightarrow$

onde o subscrito "a" é para indicar "variância agrupada" ("pooled").

Assim, o intervalo de confiança a 95% nos é dado por

com graus de liberdade.

Exemplo:

Dados n1=25 X1=60,0 S1=12
n2=15 X2=68,0 S1=10

Suponha-se que 1=2=x.
Achar o intervalo de confiança para 1-2 a 95%

$$Sa2=25122+1510225+15-2=3600+150040-2=510038=134,21$$

$$Sa=134,21=11,58$$

$t0,025\sim2,021$ para $v=25+15-2=38$ g.l.

Logo

$$1-2=60,0-68,0\pm2,02111,58125+115$$

$$=-8,0\pm2,02111,580,33$$

$$-8,0\pm7,6$$

R:

Consideremos agora o caso da diferença na taxa de crescimento do PNB (65-80) entre países socialistas e Não-socialistas. O Output produzido pelo SPSS foi o seguinte:

Independent samples of VAR23 PAIS SOCIALISTA=1

Group 1: VAR23 EQ 0 Group 2: VAR23 EQ 1

t-test for: VAR24 TAXA CRESCIMENTO A.A. PNB-65-80

	Number of Cases	Mean	Standard Deviation	Standard Error
Group 1	113	2.4558	2.283	.215
Group 2	17	1.1353	1.511	.367

Pooled Variance Estimate				Separate Variance Estimate		
F Value	2-Tail Prob.	t Value	Degrees of Freedom	2-Tail Prob.	t Value	Degrees of Freedom	2-Tail Prob.
2.28	.062	2.31	128	.023	3.11	28.40	.004

Calculando as estatísticas obtemos:

$$Sa2=1132,2832+171,13532113+17-2=610,8775128=4,7725$$

$$Sa=4,7725=2,1846$$

$$t=X1\text{-}X2Sa1n1+1n2$$

$t0,025=1,980$ para $v=113+17-2=128\sim120$g.l.

O intervalo de confiança a 95% então será

$$1\text{-}2=2,4558\text{-}1,1353\pm1,9802,18461113+11175\_0,2601$$

$$=1,3205\pm1,1252$$

$\rightarrow$ Rejeita-se H_0

Compare estes resultados com aqueles discutidos na aula anterior.

Para o teste clássico calculamos:

$$t=1,32052,18460,2601=2,31$$

Se nossa hipótese alternativa for que os países não-socialistas experimentaram taxas maiores de crescimento, então

$$H1:1>2$$

e o valor de t acima deve ser comparado com o valor de t crítico $t0,05=1,645$ para v=120g. de liberdade (aproximadamente, já que v aqui é de 128). Como o t é maior que o *t* crítico, então a decisão a ser tomada é a de rejeitar a hipótese nula.

O que aconteceria se usássemos um teste de 1%?

Importante: Observe que o valor de *t* obtido é precisamente aquele que no output do SPSS vem identificado como "*Pooled Variance Estimate*"

T value	Degrees of Freedom	2-tail Prob.
2.3'	128	.023

Assim, para fazer o "Teste de *t*" pelo SPSS basta consultar o valor de *t* no output sob o título de "*Pooled Variance Estimate*", o teste de hipótese sendo decidido na base ou da comparação deste *t* com o valor crítico tabulado ou – bem mais fácil – verificando se o valor da probabilidade relatado está abaixo (e rejeita-se H_0) do valor crítico estabelecido para a probabilidade: 5% no caso de hipótese alternativa direcionada, 2,5% quando H_1 não é direcionado; 1% ou 0,5% no outro nível convencional.

No caso do nosso exemplo, como H_1 é direcionado, então a comparação do valor da probabilidade de 0,023 é comparado com 0,05 $\rightarrow$rejeita-se H_0.

PARTE 3: MÉTODOS PARA ANÁLISE DE ASSOCIAÇÃO

CAPÍTULO 9 – ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Qui-quadrado

A pesquisa eleitoral do Rio de Janeiro em 1978 indicou uma proporção de indivíduos se identificando com o então MDB na ordem de 75,79%. Relacionando com o nível educacional do eleitor, obteve-se a seguinte tabela:

Identificação Partidária	Nível Educacional		
	Até Primário	Secundário ou mais	Total
ARENA	36 (21,95%)	79 (25,40%)	115 (24,21%)
MDB	128 (78,05%)	232 (74,60%)	360 (75,79%)
Total	164 (100%)	311 (100%)	475 (100%)

A questão que se coloca é se existe ou não uma associação significativa entre nível educacional e identificação partidária. No caso da tabela acima existem duas maneiras alternativas de responder a esta questão:

1. A maneira que já conhecemos. Observamos que existe uma diferença percentual (proporcional) de $p1\text{-}p2=0{,}7805\text{-}0{,}7460=0{,}0345$, indicando que, aparentemente pessoas *menos* instruídas tendem a se identificar *mais* com o MDB do que as mais instruídas. Observamos também que na tabela acima *só existe uma diferença percentual*, uma vez que, dado que a soma das percentagens em cada coluna é sempre 100%, a diferença proporcional dos indivíduos se identificando com ARENA por nível educacional é a mesma que a anterior com sinal invertido (isto é, -0,0345). Isto é fácil de entender: chamamos as proporções de indivíduos se identificando com ARENA para cada nível educacional de $p1$ e $p2$.

Então, sabendo que $p1=1\text{-}p1$ e $p2=1\text{-}p2$, temos que

$$p1\text{-}\,p2=1\text{-}p1\text{-}1\text{-}p2=1\text{-}p1\text{-}1+p2=p2\text{-}p1$$

A questão que se coloca pode ser formulada nos seguintes termos: é a diferença observada $p1\text{-}p2=0{,}0345$ *significativamente diferente de zero*?

Como já vimos, posso testar essa proposição de maneira clássica estabelecendo

$$H0:\ 1\text{-}2=0$$

Agora, isso equivale a dizer $1=2=\pi$, ou seja, ambas as populações tem uma proporção π de pessoas se identificando com o MDB. Sabemos também que a distribuição amostral de $p1\text{-}p2$ tem média $1\text{-}2$e desvio-padrão$1(1\text{-}1)n1+2(1\text{-}2)n2$, ou seja

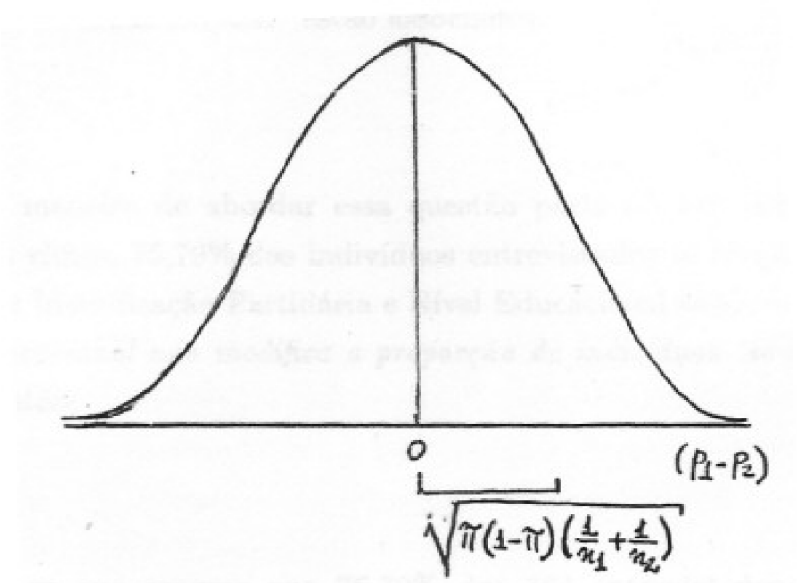
$$p1\text{-}p2\sim N1\text{-}2,\ 1(1\text{-}1)n1+2(1\text{-}2)n2$$

O teste clássico de hipótese implica em calcular a probabilidade de ocorrência da $p1\text{-}p2$ observada na suposição de que H_0 é verdadeira. Mas, se H_0 é verdadeira, então $1\text{-}2=0$ e, sendo $1=2=\pi$, tenho que o desvio padrão da distribuição amostral é dado por

$$p1\text{-}p2 = \pi(1\text{-}\pi)n1 + \pi(1\text{-}\pi)n2 = \pi(1\text{-}\pi)1n1 + 1n2$$

Ou seja, se H_0 é verdadeira,

$$p1\text{-}p2 \sim N0, \pi(1\text{-}\pi)1n1 + 1n2$$



O cálculo da probabilidade de ocorrência de um valor p1-p2 ou algo ainda maior, sendo H_0 verdadeiro, é, portanto simples. Acontece que desconhecemos o valor de π, isto é a proporção na população que se identifica com o MDB. Como estimá-lo? Voltando ao início, sabemos que a proporção em toda a amostra se identificando com o MDB é p=0,7579. Então, o melhor que podemos fazer é estimar π através de p, e assim, podemos calcular a probabilidade sob a curva normal através de (lembrando que n1=164 e n2=311)

$$Z = p1\text{-}p2\text{-}1\text{-}2p(1\text{-}p)1n1 + 1n2 = 0{,}03450{,}7579(1\text{-}0{,}7579)1164 + 1311$$

$$= 0{,}835 \sim 0{,}84$$

Obtendo

$$\text{Prob } Z \geq 0{,}84 = 0{,}20 \quad \text{ou} \quad 20\%$$

Como esse valor é muito inferior a qualquer valor crítico convencional, não podemos rejeitar H_0. Ou seja, a diferença observada não é significativamente maior que zero a qualquer nível convencional. Alternativamente, rejeito a idéia de que "identificação partidária" e "nível educacional" estão associados.

2. A outra maneira de abordar esta questão parte do conceito de *independência*. Como já vimos 75,79% dos indivíduos entrevistados se identificam com o MDB. Então

se “identificação partidária” e “nível educacional” são independentes (ou seja, *nível educacional não modifica a proporção de indivíduos se identificando com o MDB*) então:

a. Temos que esperar que 75,79% dos 164 entrevistados com educação até primário se identifiquem com o MDB. Ou seja, 0,7579 x 164 = 124,29 indivíduos.

b. Temos que esperar que 75,79% dos 311 entrevistados com educação secundário ou mais se identifiquem com o MDB. Ou seja 0,7579 x 311 = 235,71 indivíduos.

c. Como o total de indivíduos com educação “até primário” é 164 e o total com educação “secundário ou mais” é 311, tenho que esperar, no caso das duas variáveis serem independentes, que 164-124,30 = 39,71 indivíduos com educação até primário e 311-235,71 = 75,29 indivíduos com educação secundário ou mais se identifiquem com ARENA.

Podemos, pois, montar uma tabela com as freqüências esperadas *no caso da hipótese de que as variáveis são independentes ser verdadeira*.

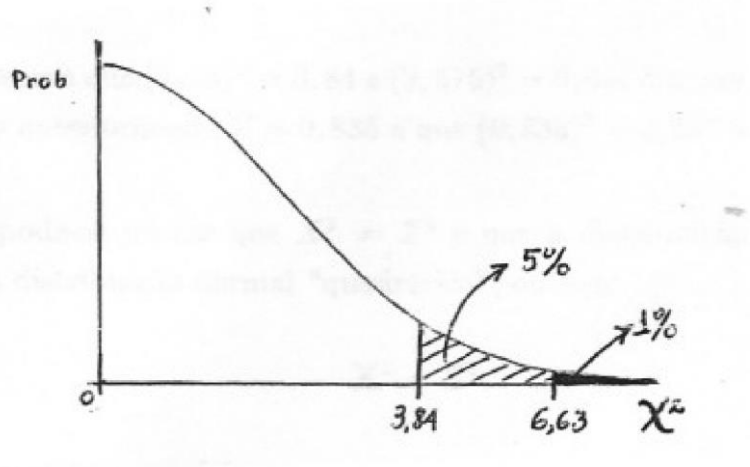
Identificação Partidária	Nível Educacional	
	Até Primário	Secundário ou mais
ARENA	39,71	75,29
MDB	124,29	235,71
Total	164	311

Podemos então comparar as duas tabelas, a tabela com os valores *observados* e a tabela com os valores *esperados* no caso das variáveis serem independentes. É claro que a diferença fo-fe, frequencia observada menos frequencia esperada, indica quanto os dados observados se *desviam* da independência, quanto maior for esse desvio fo-fe, maior será a associação entre as variáveis. Assim, se somarmos todos os desvios fo-fe, teremos o quanto os dados observados se desviam agregadamente da independência. No entanto, pela mesma razão pela qual devemos elevar a diferença ao quadrado no caso do cálculo da variância, teremos aqui que calcular fo-fe2. Assim, essa expressão mede a magnitude dos desvios totais entre o observado e o esperado no caso de independência. É claro que essa expressão será tanto maior quanto maior for o tamanho da amostra que estivermos trabalhando. Seria conveniente trabalhar com *desvios relativos ao esperado*, e não com desvios totais (absolutos). Dessa forma, podemos definir uma estatística, que chamaremos X^2

$$X2=fo-fe2fe$$

que representa a soma (o quantum agregado) dos desvios relativos *quadráticos* entre o observado e o esperado, relativamente ao esperado no caso de independência.

Agora, pode-se demonstrar que, *sendo a hipótese nula de independência entre variáveis verdadeira na população*, a estatística X^2 se distribui como a curva teórica de probabilidades 2 (qui-quadrado), que (no caso específico da tabela acima, em que só temos 1 diferença proporcional, como já vimos) tem mais ou menos a seguinte forma



Observe que valores de $2 \geq 3{,}84$ têm uma probabilidade de 5% de ocorrer e $2 \geq 6{,}63$ tem uma probabilidade de 1% de ocorrer, *sendo a hipótese de independência verdadeira*. Podemos, pois, calcular o valor da *estatística* X^2 e compará-la com os valores críticos de 2 escolhidos. Se o valor de X^2 for maior que 3,84, então trata-se de um desvio com probabilidade de ocorrer sendo a hipótese nula verdadeira menor que 5%. Isso me levaria a rejeitar a hipótese nula de independência, dizendo que as variáveis estão associadas. Caso contrário, se X^2 for menor que 3,84 então não posso rejeitar a hipótese nula ao nível de 5%, e direi que as variáveis não estão significativamente associadas.

Calculando então para os dados da nossa tabela, temos:

$$x2=36\text{-}39{,}71239{,}71+79\text{-}75{,}29275{,}29+128\text{-}124{,}292124{,}29+232\text{-}235{,}712235{,}71=$$

$$=0{,}3457+0{,}1828+0{,}1107+0{,}0582=0{,}697$$

Nota: cálculo feito com mais dígitos significativos do que o indicado.

Como o valor de X^2 está muito abaixo do valor crítico de 5%, isto é, 3,84, *não podemos rejeitar a hipótese nula de independência* a nenhum nível convencional. Assim, novamente, concluímos que identificação partidária e nível educacional *não* estão associados.

Agora observe o seguinte: sabemos que numa distribuição Normal $\text{prob } Z \geq 1{,}96 = 0{,}05$ ou 5%; que $\text{prob } Z \geq 2{,}575 = 0{,}01$ ou 1%.

Agora observe que $1{,}962 = 3{,}84$ e $2{,}5752 = 6{,}63$. Da mesma maneira, observe que calculamos anteriormente $Z = 0{,}835$ e que $0{,}8352 = 0{,}697 = X2$.

De fato, pode-se provar que $2 = Z2$ e que a distribuição de Qui-quadrado é idêntica a uma distribuição normal "quadrática", ou seja

$$2 = Z2$$

no caso em que só temos 1 diferença proporcional na tabela.

Assim, o teste de diferença de proporções e o teste de qui-quadrado *são precisamente a mesma coisa*. Na verdade é o mesmo teste só que feito numa curva Normal cujos valores foram elevados ao quadrado (sendo, portanto, todos *positivos*).

Vejamos agora os dados equivalentes para a pesquisa eleitoral do Rio de Janeiro em 1982

Identificação Partidária	Nível Educacional		
	Até Primário	Secundário ou mais	Total
PDS	45 (24,1%)	86 (26,7%)	131 (25,7%)
PMDB	45 (24,1%)	49 (15,2%)	94 (18,5%)
Outros (PDT, PTB e PT)	97 (51,9%)	187 (58,1%)	284 (55,8%)
Total	187	322	509

Observe que agora temos duas diferenças proporcionais, entre colunas, a terceira sendo automaticamente determinada pelo fato de que a soma das proporções em cada coluna é necessariamente 1.

A consequência do fato de termos 2 diferenças proporcionais é que o teste de diferenças visto anteriormente não mais pode ser aplicado *para concluir* sobre a relação entre as variáveis *em geral* (embora possa aplicá-lo para examinar uma diferença proporcional específica em que estejamos interessados). O teste do Qui-quadrado, no entanto, nos proporciona uma generalização do teste de diferenças, testando a existência *agregadamente* de diferenças significativas dentro da tabela.

Vejamos a aplicação do teste de qui-quadrado a essa tabela de 1982.

1. Cálculo das frequências esperadas no caso de independência – procedimento idêntico ao visto anteriormente

Identificação Partidária	Nível Educacional	
	Até Primário	Secundário ou mais
PDS	48,1	82,9
PMDB	34,5	59,5
Outros (PDT, PTB e PT)	104,4	179,6
Total	187	322

2. Cálculo da estatística X^2:

$$x2=45\text{-}48{,}1248{,}1+86\text{-}82{,}9282{,}9+45\text{-}34{,}5234{,}5+49\text{-}59{,}5259{,}5+97\text{-}104{,}42104{,}4+187\text{-}179{,}62179{,}6=$$

$$=0{,}20+0{,}12+3{,}2+1{,}85+0{,}52+0{,}30=6{,}19$$

3. Cálculo do ponto crítico de 2 (a 5% ou 1% ou qualquer nível que se queira) para comparar com o valor de X^2

Lembre-se que no caso da tabela de 78, mostramos que 2=Z2. Isso porque só tínhamos 1 diferença percentual *livre* na tabela, e assim, o teste de X^2 é idêntico ao teste de diferença de proporções. Mas agora temos 2 proporções livres e, portanto, não podemos utilizar a mesma curva de 2.

Então, precisamos saber que a distribuição de 2 é, na verdade, uma soma de distribuições normais padrão ao quadrado

$$2=Z12+Z22+...+Zv2$$

sendo *v* o número de *diferenças proporcionais livres* na tabela. Em outras palavras, *existe uma curva de* 2 *diferente para cada número de diferenças livres possível*. No nosso caso, como temos 2 diferenças livres, tenho que pesquisar a distribuição para

$$2=Z12+Z22$$

Importante: O número de proporções livres na tabela recebe a denominação de *graus de liberdade*, e para uma tabela qualquer, podemos calcular o número de graus de liberdade através da expressão

$$g.l=l-1c-1$$

onde l é o número de *linhas* na tabela e c é o número de *colunas*. No caso da tabela de 1982, l=3 e c=2, então:

$$g.l=3-12-1=2\times1=2 \text{ graus de liberdade}$$

Pesquisando na curva correspondente de 2 com 2 graus de liberdade (ver tabela anexa), encontro os valores críticos de 2 a 5% e a 1%:

$$5\%2=5,99 \quad \text{e} \quad 1\%2=9,21$$

Se utilizarmos o nível de significância de 5%, descobrimos que a estatística $X^2=6,19$ é ligeiramente maior que o valor crítico (5,99), implicando numa probabilidade de ocorrência, sendo a hipótese de independência entre as variáveis verdadeira, inferior a 5%. Rejeito, pois a hipótese nula de independência, e digo que as variáveis estão significantemente associadas.

O teste de Qui-quadrado generaliza, portanto, o teste de diferença proporcional, permitindo afirmar a existência ou não de diferenças significativas (agregadamente) entre as v diferenças proporcionais na tabela.

Vamos a um último exemplo numérico, tirado do artigo de Amaury de Souza "Exposição aos Meios de Comunicação de Massa no Rio de Janeiro: um estudo preliminar" (*Dados*, 4 (1968): 145-168).

Exposição à Televisão				
Ocupação	Alta	Média	Baixa	Total
A. Prof. Liberais, Diretores e Proprietários	25	12	3	40 (7,31%)
B. Trabalhadores Não-manuais e prof. de padrão inferior	132	55	23	210 (38,39%)
C. Trabalhadores manuais especializados	92	49	39	180 (32,91%)
D. Trabalhadores Manuais Não-especializados	30	34	53	117

				(21,39%)
Total	279	150	118	547 (100%)

1. Cálculo das frequências esperadas no caso de independência:

Ocupação	Alta	Média	Baixa
A.	20,4	11	8,6
B.	107,1	57,6	45,3
C.	91,8	49,4	38,8
D.	59,7	32,0	25,2

2. Cálculo da estatística X^2

Ocupação	Exposição	Frequencia	F esperada	fo-fefe2
A	A	25	20,4	1,04
A	M	12	11	0,91
A	B	3	8,6	3,65
B	A	132	107,1	5,79
B	M	55	57,6	0,12
B	B	23	45,3	10,98
C	A	92	91,8	0,00
C	M	49	49,4	0,00
C	B	39	38,8	0,00
D	A	30	59,7	14,78
D	M	34	32	0,13
D	B	53	25,2	30,67
TOTAL				68,07

$$x2=68,07$$

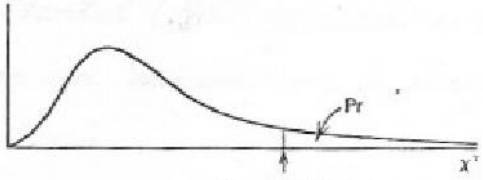
A tabela tem l=4 e c=3. Portanto temos

$$g.l=4\text{-}13\text{-}1=3\times2=6 \text{ graus de liberdade}$$

3. Pesquisando na tabela de 2 com 6 graus de liberdade encontro o valor crítico a 1% de 16,8. Como o valor encontrado para a estatística X^2 ultrapassa de muito o valor crítico

de 1%, rejeito a hipótese nula de independência entre as variáveis "Exposição à Televisão" e "Ocupação" a qualquer nível convencional.

Pontos Críticos de χ^2



Ponto crítico

g.l. \ Pr	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	19.4	23.5	26.3	28.8	32.0	34.3	39.3
17	20.5	24.8	27.6	30.2	33.4	35.7	40.8
18	21.6	26.0	28.9	31.5	34.8	37.2	42.3
19	22.7	27.2	30.1	32.9	36.2	38.6	32.8
20	23.8	28.4	31.4	34.2	37.6	40.0	45.3
21	24.9	29.6	32.7	35.5	38.9	41.4	46.8
22	26.0	30.8	33.9	36.8	40.3	42.8	48.3
23	27.1	32.0	35.2	38.1	41.6	44.2	49.7
24	28.2	33.2	36.4	39.4	32.0	45.6	51.2
25	29.3	34.4	37.7	40.6	44.3	46.9	52.6
26	30.4	35.6	38.9	41.9	45.6	48.3	54.1
27	31.5	36.7	40.1	43.2	47.0	49.6	55.5
28	32.6	37.9	41.3	44.5	48.3	51.0	56.9
29	33.7	39.1	42.6	45.7	49.6	52.3	58.3
30	34.8	40.3	43.8	47.0	50.9	53.7	59.7
40	45.6	51.8	55.8	59.3	63.7	66.8	73.4
50	56.3	63.2	67.5	71.4	76.2	79.5	86.7
60	67.0	74.4	79.1	83.3	88.4	92.0	99.6
70	77.6	85.5	90.5	95.0	100	104	112
80	88.1	96.6	102	107	112	116	125
90	98.6	108	113	118	124	128	137
100	109	118	124	130	136	140	149

Para interpolar cuidadosamente, ver Tab. X.

Exercícios:

1. Analisar a seguinte tabela extraída de um estudo clássico classe social e doença mental (Hollingshead e Realich, 1958) em que se relaciona o disgnóstico da doença com a terapia que foi indicada:

	Cor			
Diagnóstico: Tipo de Psicose	Psicoterapia	Terapia Orgânica	Custódia	Total
Afetivas	30	102	28	160
Alcoólicas	48	23	20	91
Orgânicas	19	80	75	174
Esquizofrênicas	121	344	382	847
Senis	18	11	141	170
	236	560	646	1442

2. Na pesquisa sobre identidades coletivas realizada pelo IDESP em São Paulo (capital) em 1986, obtivemos a seguinte tabela relacionando a Cor do Entrevistado (atribuída pelo entrevistador) e sua opinião sobre se Existe ou não discriminação contra Pretos e Mulatos no Brasil.

	Cor			
Opinião	Branca	Preta	Parda	Amarela
Existe Discriminação	272	22	76	14
As Oportunidades são iguais	132	11	38	1

O que você concluiria? Justifique.

3. Analise a seguinte tabela, oriunda da pesquisa eleitoral de 1982 no Rio de Janeiro, em que se relaciona a Opinião sobre Controle Governamental dos Sindicatos com a Condição de Sindicalização:

	Opinião: O Governo			
Sindicalização	Controla	Não Controla	Não sabe	
Sindicalizado	80	28	19	127

Não sindicalizado	172	96	206	474
	252	124	225	601

4. A Pesquisa Nacional de Saúde e Nutrição, feita pelo IBGE em 1989, permite comparar o estado nutricional das crianças menores de 5 anos com o de suas mães. A avaliação antropométrica da criança foi feita da seguinte forma: a criança "normal" não apresenta déficit de peso nem de estatura; a criança com crescimento "deficiente" apresenta um desses défitis ou ambos, segundo a classificação de Waterlow. A avaliação antropométrica da mãe foi feita da seguinte forma: "magra" apresenta Índice de Massa Corporal inferior a 20; "normal", entre 20 e 25, inclusive, e com "sobrepeso" ou "obesa" com Índice superior a 25. Os resultados são os seguintes:

	Mãe			
Crianças	Magra	Normal	Gorda	
Normal	1938	6784	2735	11457
Com déficit	307	673	184	1164
	2245	7457	2919	12621

Analisando estes dados, o que você pode dizer a respeito da fome no Brasil?

CAPÍTULO 10 – ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS E QUANTITATIVAS (ANOVA) (REPETE CAPÍTULO 2 ATÉ PÁGINA 4)

Análise de Variância

O tópico "Análise de Variância" encobre uma gama de modelos para análise de dados, modelos esses que envolvem uma variável dependente *Y* mensurada a nível de intervalo ou razão, e uma ou mais variáveis independentes, escolhidas para explicar *Y*, mensuradas ao nível nominal. Em outras palavras, para cada categoria (ou combinação de categorias) da minha variável independente (ou variáveis independentes), fazemos um conjunto de observações sobre a variável dependente. O propósito da Análise de Variância é descobrir se existem diferenças no nível (médio) da variável dependente *Y* entre as categorias das variáveis independentes.

Vamos estar aqui interessados no caso em que apenas uma variável nominal está envolvida, ou seja, na Análise da Variância a Um Fator. Como a variável nominal tem *k* categorias, o nosso problema será portanto o de descobrir se esses *k* grupos definidos pelas categorias da variável dependente se caracterizam por níveis (isto é, médias) diferentes na variável dependente *Y*.

Nota: Quando *k*=2 a comparação das médias entre os grupos (categorias) é feita por um teste de hipóteses específico envolvendo a distribuição "t" de Student.

Assim, os dados para a Análise da Variância podem ser representados na forma

Categoria *i*

i=1	*i*=2	...	*i*=*k*
Y11	Y11	...	Y11
Y12	Y11	...	Y11
...	...		...
Y1n1	Y1n2	...	Y1nk

Observe-se que para cada um dos *k* grupos temos ni (*i*=1, 2, ..., *k*) observações diferentes.

Modelo a 1-Fator

Supõe-se que os dados para a Análise da Variância a 1 fator foram gerados por um modelo (na População) que pode ser descrito como

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

i = 1, 2, ..., *k*
j = 1, 2, ..., ni

onde Yij é o valor de *Y* na jésima categoria. Yij foi produzido pela soma de 3 componentes:

μ = nível geral (média) de *Y*

i= efeito sobre *Y* de pertencer à categoria *i*

eij= efeito sobre *Y* de outras variáveis que não sejam "pertencer a categoria *i*", que se supõe não relevantes. Também chamada de "variável residual".

Aqui, para podermos estimar os diversos parâmetros (e fazer os testes de hipóteses), precisamos fazer hipóteses quanto à natureza dessa "variável residual". Mais especificamente suporemos que

$$e_{ij} \sim N(0, \sigma_e^2)$$

Os verdadeiros valores dos parâmetros μ, 1, 2, ..., k e do termo residual são naturalmente desconhecidos, sua estimação sendo o objeto próprio (bem como o teste de significância de suas eventuais diferenças) da Análise de Variância.

Para fins de exposição, entretanto, suporemos que para um caso específico em que *k* =3 categorias os parâmetros na população são os seguintes

μ=10; 1=4; 2=1; 3=-5

Observe-se que como μ é o nível (média) geral da variável *Y*, então necessariamente

$$\sum_{i=1}^{k} \alpha_1 = 0$$

Conhecidos esses parâmetros e na ausência de efeitos da variável residual, os valores observados seriam (para, por exemplo, categorias com cinco observações cada uma) numa dada amostra obtida daquela população Yij=μ+i

Observação	Categoria 1 1=4	Categoria 2 2=1	Categoria 3 3=-5
1	14	11	5
2	14	11	5
3	14	11	5
4	14	11	5
5	14	11	5

∴ Y=10

A Variação Total entre as 15 observações pode ser medida pela soma dos quadrados

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}\right)^2$$

= 210

Como não existe variação dentro de cada grupo (dado que eliminamos o termo eij), essa Variação Total é completamente originada pelas diferenças entre grupos.

Suponhamos agora que os valores da variável residual eij são os seguintes naquela amostra

Grupo (categoria *i*)

1	2	3
-2,5	-3,3	-4
0,1	-3,5	-3,5
3,9	-0,3	0
2,0	0,4	1,0
4,9	2,4	0,4

Assim os valores observados na amostra são os seguintes

Grupo (categoria *i*)

1	2	3	
11,5	7,7	1	Y=9,87
14,1	7,5	1,5	Y1=15,68
17,9	10,7	5	Y2=10,14
16,0	11,4	6	Y3=3,78
18,9	13,4	5,4	

A Variação Total agora é

$$i=1kj=1niYij-Y2=437,38$$

devida agora a uma combinação de variação intra e inter-grupos.

Análise de Variância: Estimativas

Os parâmetros μ, α1,2,..., k são estimados por , 1,2,..., k.

A estimativa de Mínimos Quadrados implica em minimizar a diferença entre o valor observado e o valor predito pelo modelo.

$$Yij=+i$$

Ou seja, minimiza-se

$$Q=ikjniYij-Yij2=ikjnieij2=ikjniYij--i2$$

sujeito à restrição que

$$iki=0$$

Dessa minimização resultam os seguintes estimadores

$$\hat{\mu} = \bar{Y}\hat{\alpha}_i = \bar{Y}_i - \bar{Y}$$

e

Assim, para o nosso exemplo, as estimativas dos parâmetros seriam

$$=Y=9,87$$

$$1=Y1-Y=15,68-9,87=5,81$$

$$2=Y2-Y=10,14-9,87=0,27$$

$$3=Y3-Y=3,78-9,87=-6,09$$

os quais, naturalmente, discrepam dos valores na População.

Teste das Diferenças Entre Grupos

Observe-se que o desvio para cada observação na amostra pode ser escrito como

$$\underbrace{\left(Y_{ij} - \bar{Y}\right)}_{desvio\ Total} = \underbrace{\left(Y_{ij} - \bar{Y}_i\right)}_{desvio\ DENTRO\ do\ Grupo} + \underbrace{\left(\bar{Y}_i - \bar{Y}\right)}_{desvio\ ENTRE\ os\ grupos}$$

Se agora quadramos esse desvio e somamos ao longo de todas as observações

$$i=1kj=1niYij-Y2=i=1kj=1niYij-Yi+Yi-Y2=$$

$$=i=1kj=1niYij-Y2+i=1kj=1niYij-Yi2+2i=1kj=1niYij-Yi-Yi-Y$$

Agora, o último termo à direita da equação é

2i=1kj=1niYij-Yi-Yi-Y=2i=1kYij-Yj=1niYij-Yi⏟=0=0

uma vez que Yi-Y é constante para cada valor de *i* e j=1niYij-Yi=0 necessariamente por definição de Yi

Logo

i=1kj=1niYij-Y2=i=1kj=1niYij-Yi2+i=1kj=1niYi-Y2⏟ni vezes

Como no último termo Yi-Y é constante para cada valor de *i*, temos que j=1niYij-Y2=niYi-Y2.

Então

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2 + \sum_{i=1}^{k} n_i\left(\bar{Y}_i-\bar{Y}\right)^2$$

Soma dos Quadrados Total (TSS) | Soma dos Quadrados Intra-Grupos (WSS) | Soma dos Quadrados Inter-Grupos (BSS)

Variância Total | "Variação não-explicada" pela variável categórica | "Variação explicada" pela variável categórica

Ou seja, podemos decompor a Variação Total (Soma dos Quadrados) em 2 componentes: Variação Inter e Intra-Grupos.

A Variação Inter-Grupos (BSS) é calculada com (k-1) graus de liberdade.
A Variação Intra-Grupos (WSS) é calculada com (n-k) graus de liberdade, onde n=ini.
Então a Variação Total (TSS) é calculada com (k-1) + (n-k) = (n-1) graus de liberdade.

Dividindo cada valor por seus graus de liberdade, obtemos as chamadas "Somas dos Quadrados Médias"

BMS (Soma dos Quadrados Média Inter-Grupos) = BSS/(k-1)
WMS (Soma dos Quadrados Média Intra-Grupos) = WSS/(n-k)

A razão

F=BMSWMS=variação "explicada"variação "não-explicada"=ikYi-Y2k-1i=1kj=1niYij-Y2n-k

se distribui de acordo com a distribuição F de Snedecor com (k-1) e (n-k) graus de liberdade no caso em que a hipótese nula

H0: 1=2=...=k=0

é verdadeira.

Estas informações nos permitem construir a seguinte tabela sumarizadora da Análise de Variância, a chamada tabela ANOVA:

ANOVA: Tabela da Análise de Variância a 1 Fator

Fonte	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	Razão-F
Entre-Grupos	iniYi-Y2=BSS	*(k-1)*	BMS=BSS(k-1)	BMS/WMS
Intra-Grupos	ijYij-Yi2=WSS	*(n-k)*	WMS=BSS(n-k)	
Total	ijYij-Y2=TSS	*(n-1)*		

Obviamente, quanto maior for a variação entre grupos relativamente à variação intra-grupos, maior a significância das diferenças de médias entre grupos. A razão F mede o quão significante é essa diferença, possibilitando o seu exame probabilístico.

Construindo a tabela ANOVA para os dados do nosso exemplo, obtemos:

Fonte	S. S	g. l.	MS	F
Entre-Grupos	354,59	2	177,30	25,70
Intra-Grupos	82,79	12	6,90	
Total	437,38	14		

A análise dos dados exemplificada acima foi feita em duas etapas: a primeira consistiu em estimar os parâmetros μ, 1, 2, ..., k. As estimativas 1, 2, ..., k indicam a extensão em que as médias dos grupos diferem entre si.

A segunda etapa consistiu na comparação entre as médias dos grupos, comparação essa que envolveu o cálculo da variação (soma dos quadrados) entre-grupos como uma medida do quanto agregadamente esses grupos diferem entre si. Essa medida é por sua vez comparada com a variação restante, intra-grupos, utilizando-se para isso a Razão-F. Claramente, quanto maior a variação "explicada" relativamente à variação "não-explicada" (Inter/Intra), maior a significação das diferenças entre os grupos na composição da variação Total.

Comparando-se a Razão-F empiricamente obtida com o valor tabulado para a distribuição de F na hipótese de que

$$H0: 1=2=...=k=0$$

com (k-1, n-k) graus de liberdade, podemos testar a significância das diferenças entre as médias observadas.

Em nosso exemplo, $F_{(2,12)}=25,70$. Com 2 e 12 graus de liberdade (no numerador e no denominador, respectivamente) a probabilidade de se obter um valor maior que 8,51 quando $1=2=3=0$ é de apenas 0,005. O valor encontrado é portanto altamente significativo, implicando que as diferenças entre as médias encontradas é significativamente diferente de zero. Rejeita-se, portanto, a hipótese nula de que na População

H0: 1=2=...=k=0.

CAPÍTULO 11 – REGRESSÃO LINEAR SIMPLES E MÚLTIPLA