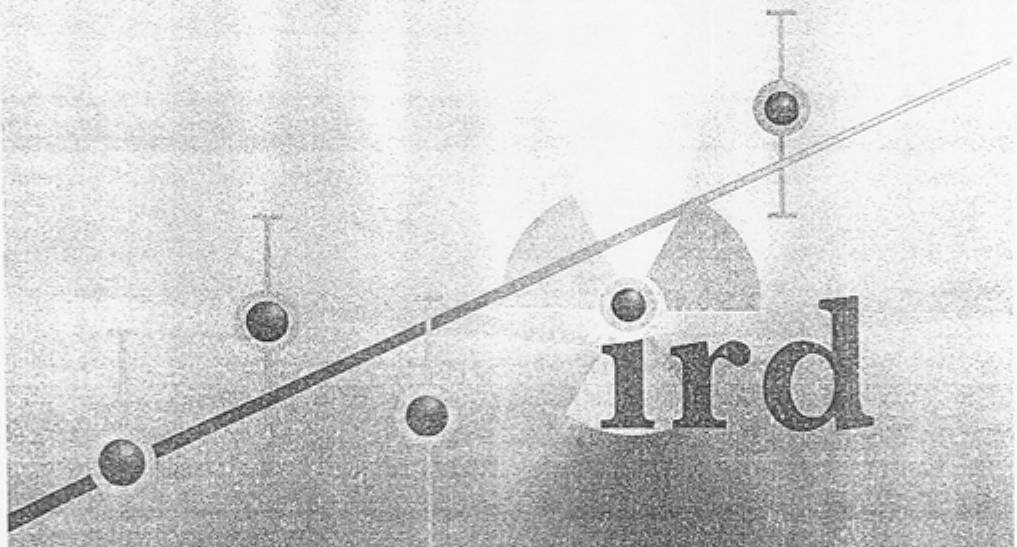
ESTATÍSTICA, TEORIA DE ERROS
E
PROCESSAMENTO DE DADOS



Elizabeth Santos de Almeida

Luiz Tauhata

LABORATÓRIO NACIONAL DE METROLOGIA
DAS RADIAÇÕES IONIZANTES

IRD – CNEN

ESTATÍSTICA E TEORIA DE ERROS
PARA RADIOPROTEÇÃO

Luiz Tauhata
Elizabeth Santos de Almeida

INSTITUTO DE RADIOPROTEÇÃO E DOSIMETRIA - CNEN

As informações contidas nesta apostila são de responsabilidade exclusiva dos autores.

APRESENTAÇÃO

Os conceitos e procedimentos apresentados nesta apostila visam, unicamente, auxiliar estudantes e colegas de trabalho, no tratamento de dados coletados em pesquisas na área de Proteção Radiológica. Representa uma coletânea de temas que aparecem com frequência e que, muitas vezes, não se tem um enfoque incial para o processamento dos dados de trabalhos de importância científica.

O texto visa resolver problemas práticos, utilizando alguns conceitos básicos de estatística. Em algumas situações, carece de maior rigor e embasamento estatístico. Mas, pela nossa experiência, acreditamos que será de utilidade para muitas pessoas.

Os autores

ESTATÍSTICA E TEORIA DE ERROS PARA RADIOPROTEÇÃO

ÍNDICE

ESTATÍSTICA, TEORIA DE ERROS E PROCESSAMENTO DE DADOS

I - INTRODUÇÃO

> " Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."
>
> - H.G.Wells -

Em sua interpretação moderna, a estatística é uma teoria e uma metodologia de deduzir e tomar decisões sob condições de incerteza. Trata também de como essas deduções podem ser estendidas além do conjunto particular de dados examinados, e de como as decisões podem ser neles baseadas, mediante apropriada análise.

Conceitos e métodos estatísticos fornecem um meio lógico,objetivo e sistemático de se tomar decisões em problemas de qualquer natureza.

Em muitas áreas da atividade humana, a tomada de decisões é baseada somente em julgamento intuitivo. Em outras, elas requerem cuidadosas pesquisas, com refinada triagem dos dados e detalhada análise estatística.

Os procedimentos estatísticos, entretanto, não substituem ou descartam a intuição e o senso comum. Ao contrário, ajudam na estruturação e condução do problema até a decisão.

A necessidade de se tratar um fenômeno estatísticamente,reside em um dos seguintes 3 pontos, ou mesmo nos 3 simultaneamente, dependendo da situação:

1. o desconhecimento do comportamento do fenômeno ou situação estudados;
2. a impossibilidade de se ter um conjunto completo de dados;

3. a impossibilidade ou dificuldade de se avaliar diretamente determinadas grandezas ou suas correlações com demais outras.

Além destas limitações, os resultados de medidas,em uma ou num conjunto de amostras, não apresentam uma transparência ou uma interpretação fácil. A adequada classificação, representação gráfica e avaliação das flutuações associadas explicitam e sugerem melhor, modelos para a descrição do fenômeno, caminhos para assegurar conclusões.

É de praxe considerar a análise estatística como uma das etapas finais de um estudo. Obviamente que isso é falacioso, pois o simples processamento de dados não pode "fazer milagres". É necessário a previsão para garantir conclusões. Dependendo do objetivo da pesquisa, a análise estatística deve anteceder a qualquer etapa. Por exemplo, na avaliação comparativa do nível médio de radioatividade de um sítio de uma instalação nuclear, antes e após operação, é imperativo o estabelecimento de um programa de coleta de amostras, com frequência, tamanho, indicadores e locais pré-estabelecidos, monitores definidos para cada nível de radioatividade e tipo de radiação a medir, fixação dos parâmetros indicativos da situação de medida como, grau de umidade, salinidade, temperatura, acidez, etc.

Muitos estudos com procedimentos estatísticos incorporados podem conduzir a conclusões temerárias ou polarizadas por interesses devido a "tendências"(bias) no conjunto de dados. Estas podem originar-se no detrimento nas respostas às questões do levantamento estatístico, na escolha da fonte de amostragem.

Existe também a influência das aproximações necessárias ao desenvolvimento e à interpretação dos estudos, fora limitações e armadilhas. Assim, um certo ceticismo com respeito aos resultados de qualquer investigação é saudável, por vezes, essencial.

II - INVESTIGAÇÃO ESTATÍSTICA E DADOS

> " Collecting data is much like collecting garbage. You must know in advance what you are going to do with the stuff before you collect it."
>
> - Mark Twain -

II.1 - POPULAÇÃO E AMOSTRA

A coleção total de ítens ou elementos que estão dentro do domínio da investigação, denomina-se "População". Podem ser indivíduos, famílias, núcleos radioativos, bioindicadores. Sua caracterização é que estabelece os limites do processo de coleta de dados, das inferências e conclusões que podem ser obtidas do estudo.

Em muitos casos, a população é definida, como por exemplo, o número de habitantes de uma cidade para efeito de um levantamento demográfico, o número de pessoas ocupacionalmente expostas à radiação numa instalação. Em outros, ela consiste das observações possíveis ou hipoteticamente possíveis, para um certo fenômeno, como o número de biorganismos afetados por efluente de um complexo industrial.

Frequentemente, é impraticável a inclusão dos elementos de uma população, num estudo. Recorre-se então, à enumeração parcial que simule, com determinado grau de confiança, o comportamento dela. Isto constitui uma "Amostra".

Os termos, população e amostra, são relativos. Um agregado de elementos que constitui uma população para um propósito, pode somente ser amostra de outro. Da amostra se extraem as variáveis observáveis da pesquisa. Grandezas quantificáveis. Delas serão obtidos os valores numéricos e suas flutuações.

A escolha e o dimensionamento da amostra seguem critérios dependentes dos níveis de precisão e discriminação requerido nas observações, do número de observações a serem realizadas, dos objetivos.

A amostra se insere numa programação ou estratégia, num campo de amostragem definido, deve conter os elementos pesquisados,que por sua vez, devem ser compatíveis com os métodos de medida e análise padronizados e disponíveis, minimizar os interferentes,otimizar os recursos.

Definidas as amostras representativas de uma população, enfrenta-se o problema, muitas vezes, do manuseio da massa de dados gerados nas observações. Se a pesquisa se iniciou semi-empiricamente, por condicionamento qualquer, surge imediatamente a necessidade de entender seu significado, de perceber a que critérios foi obediente. Se planejada, os meios de uma apresentação objetiva dos resultados já foram estabelecidos, a maneira de classificação e correlação dos valores, para permitir a vizualização das características e facilitar a interpretação.

II.2 - CLASSIFICAÇÃO DE DADOS

Consiste, essencialmente, de 4 passos:

a). escolha das classes;
b). classificação (ou registro) dos dados dentro das classes;
c). contagem do número de ítens em cada classe;
d). apresentação dos resultados em forma de uma tabela ou carta.

Desde que o passos (b) e (c) são triviais e o passo (d) é uma questão de habilidade e gosto,o primeiro passo necessita de detalhamento.

A escolha do número de classes e do alcance dos valores de cada uma,dependem da natureza dos dados e do propósito da pesquisa, embora os limites sejam arbitrários. Às vezes sua escolha é condicionada pela necessidade de comparação de resultados com outros traba lhos similares, cujos dados foram classificados de um modo habitual.

Algumas regras são observadas para a fixação do número de classes e seus limites:

a). o número de classes depende principalmente do número de medidas ou observações a classificar; geralmente se situa entre 5 e 20;

b). cada valor medido, ou observação, deve pertencer a somente uma classe;

c). as classes, em geral, possuem o mesmo tamanho, cujo valor deve ser adequadamente selecionado;

d). quando há necessidade de reduzir o número de classes ou quando poucos dos valores a classificar são maiores (ou menores) que a maioria, pode-se utilizar a "classe aberta" (por exemplo, maior que, menor que, etc).

O tamanho e o número de classes são interligados. A fixação de um define o outro. Por exemplo, o número de classes pode ser estimado como sendo o número inteiro mais próximo do obtido pela relação

$$k \cong \frac{H - L}{j}$$

onde,

k = número de classes
j = tamanho da classe
H = valor máximo observado
L = valor mínimo observado

Na Tabela 2.1 estão registrados 90 valores de Equivalente de Dose obtidos mensalmente por dosímetro pessoal, durante a manipulação de material radioativo por um técnico de laboratório.

Tabela 2.1 - Valores registrados de Equivalente de Dose(mSv)

0,81	0,69	0,93	0,88	0,42	0,57	1,15	0,77	0,77	0,92
0,39	0,87	0,87	1,02	1,34	0,73	0,74	0,59	1,05	1,07
0,51	0,35	0,71	0,71	0,86	0,67	0,94	0,96	1,17	1,22
0,81	0,84	0,57	0,23	0,95	1,29	1,01	0,92	0,82	0,77
0,63	0,61	0,31	1,16	1,25	1,02	0,85	0,73	0,67	0,52
0,48	0,58	0,65	0,76	0,63	1,05	0,82	0,84	0,79	1,02
1,11	0,91	0,85	1,03	1,15	0,74	0,63	1,02	0,94	0,99
0,71	0,84	1,12	0,86	0,91	0,97	0,72	0,86	1,01	0,77
0,82	0,95	0,83	0,92	0,94	0,61	0,73	0,68	0,81	0,75

Na Tabela 2.1 observa-se que:

$L = 0,23$

$H = 1,34$

Escolhendo-se um tamanho de classe j = 0,10 mSv, o número de Classes será:

$$k = \frac{H - L}{j} = \frac{1,34 - 0,23}{0,10} \cong 12$$

Para a classificação dos valores da Tabela 2.1 em 12 classes, é necessário definir os valores, inicial e final, de cada uma,ou seja, os seus limites.

O limite inferior da 1ª classe deve ter um valor de fácil manuseio e ser o mais próximo de L. Para L = 0,23, o valor 0,20 é o mais indicado. Assim os valores serão classificados segundo as classes:

0,20 - 0,30
0,30 - 0,40
....
....
1,20 - 1,30
1,30 - 1,40

Estes limites, assim definidos, são, na verdade, os denominados "Limites reais" ou "Fronteiras" de classe. Por simplicidade, serão doravante, designados por "limites".

A operação de classificar cada valor, seguindo a condição,

$$\text{Limite Inferior} \leq \text{valor} < \text{Limite Superior}$$

é denominada de "Tabulação" dos resultados, e o número de valores que pertencem a cada classe é denominado de "Frequência", ou "Frequência de Ocorrência". Isto é apresentado na Tabela 2.2.

Tabela 2.2. - Valores Tabulados do Equivalente de Dose Absorvida

CLASSE	TABULAÇÃO	FREQUÊNCIA
0,20 - 0,30	/	1
0,30 - 0,40	///	3
0,40 - 0,50	//	2
0,50 - 0,60	~~////~~ /	6
0,60 - 0,70	~~////~~ ~~////~~	10
0,70 - 0,80	~~////~~ ~~////~~ ~~////~~ /	16
0,80 - 0,90	~~////~~ ~~////~~ ~~////~~ ///	18
0,90 - 1,00	~~////~~ ~~////~~ ////	14
1,00 - 1,10	~~////~~ ~~////~~	10
1,10 - 1,20	~~////~~ /	6
1,20 - 1,30	///	3
1,30 - 1,40	/	1

Uma classe, em muitas situações, pode ser representada por um único valor, denominado "Símbolo de Classe". Um valor significativo para o símbolo de classe é o seu "Ponto Médio".Ele é obtido somando-se os limites reais da classe e dividindo-se por 2. Assim, os valores, o,25 , 0,35 , 0,45 , etc. correspondem aos pontos médios das classes: 0,20 - 0,30 ; 0,30 - 0,40 ; 0,40 - 0,50.

II.3 - REPRESENTAÇÃO GRÁFICA

Para efeito de interpretação, tabelas, como a 2.2. carecem de apelo visual. Não evidenciam as características da "Distribuição de Frequência" obtida na tabulação. Recorre-se, então, à representação gráfica. Existem alternativas várias para esta representação. Certos tipos de dados conduzem a certos tipos de gráficos.Aqui,serão descritas somente as mais importantes sob o ponto de vista da Radioproteção.

Para representar dados classificados em grupos distintos, os "Gráficos de Barra" são bastante ilustrativos. Neles somente a altura das barras é proporcional à frequência das classes. A largura das barras não tem nenhum significado. Uma simples linha seria suficiente. Uma convencional separação entre as barras é também utilizada, por motivos estéticos. Por simplicidade, a largura das barras e a separação entre elas é uniforme.

Na Fig.2.1, tem-se uma representação deste tipo. Estão representadas as 251 inspecções de equipamento de Raios-X diagnóstico feitas pelo Instituto de Radioproteção e Dosimetria em 1982, agrupadas segundo o tipo de aparelho ou técnica utilizados.

Esta mesma representação pode ser usada para expressar dois gráficos num só, oferecendo ainda a vantagem de permitir uma interpretação comparativa das frequências para o mesmo grupo de comparação. Isto pode ser visto na Fig.2.2., onde o risco, medido em termos de fatalidades/ano e normalizado por megawatt/ano, associado às diversas fontes alternativas de energia foi expresso em função dos diversos tipos, e para o público e pessoal ocupacional.

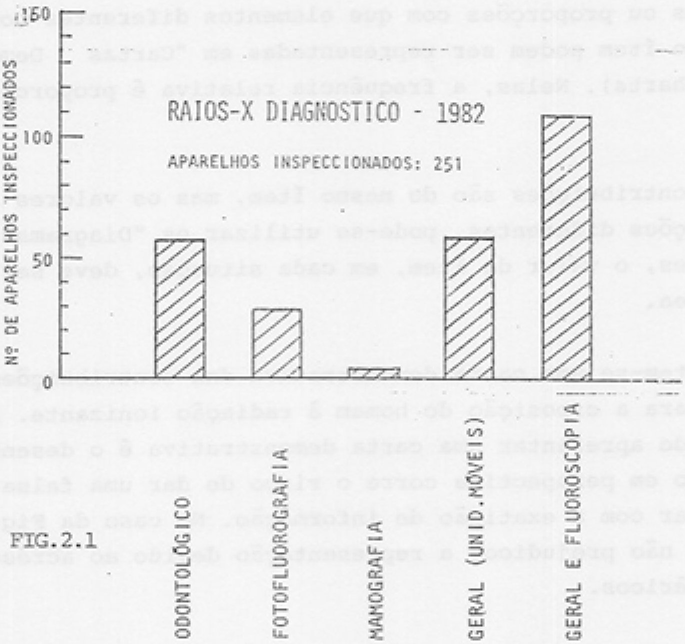
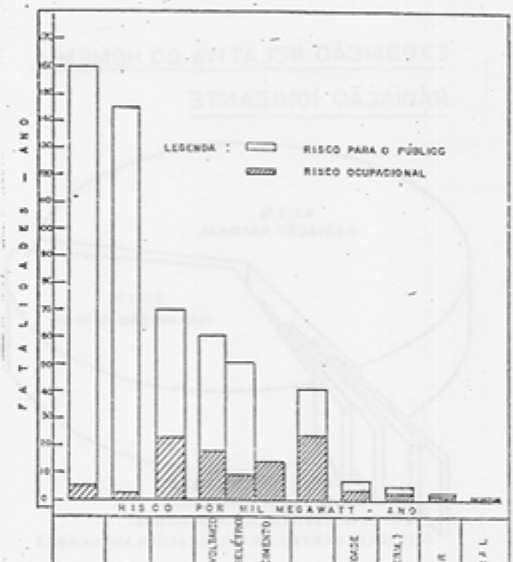


FIG.2.1

FIG.2.2



DR. HERBERT INHABER. NEW SCIENTIST, LONDON ; RISK OF ENERGY PRODUCTION, 1978 - NO AECD-1119; IEA BULLETIN - VOL. 21, Nº 1

Porcentagens ou proporções com que elementos diferentes contribuem para o mesmo ítem podem ser representadas em "Cartas Demonstrativas" (pie charts). Nelas, a frequência relativa é proporcional à área.

Quando as contribuições são do mesmo ítem, mas os valores corresdem a situações diferentes, pode-se utilizar os "Diagramas" (pictograms). Neles, o valor do ítem, em cada situação, deve ser proporcional à área.

Na Fig.2.3 tem-se uma carta demonstrativa das contribuições médias relativas para a exposição do homem à radiação ionizante. A maneira correta de apresentar uma carta demonstrativa é o desenho frontal. A visão em perspectiva corre o risco de dar uma falsa impressão ou faltar com a exatidão de informação. No caso da Fig.2.3 a perspectiva não prejudicou a representação devido ao acréscimo dos valores numéricos.

Fig.2.3

Na Fig.2.4 tem-se diagramas representativos do número de peixes de duas espécies, coletadas no período de janeiro a dezembro de 1982, agrupados segundo a estação do ano, dentro do programa de amostragem para monitoração ambiental do sítio do reator nuclear de Angra dos Reis, do IRD. Neles, as áreas são proporcionais à frequência relativa. Se fosse adotado somente o comprimento do peixe para representar a frequência relativa, ter-se-ia uma falsa impressão dos dados de coleta.

II.3.1. Histograma

Para dados apresentados em escala ou ordem numérica, utiliza-se o "Histograma", onde as áreas representam as frequências. Geralmente os intervalos de classe são iguais. Neste caso, as alturas dos retângulos de representação, são proporcionais às frequências.Consta de uma representação de medidas,ou observações,agrupadas numa escala horizontal, com as frequências de classe na escala vertical.

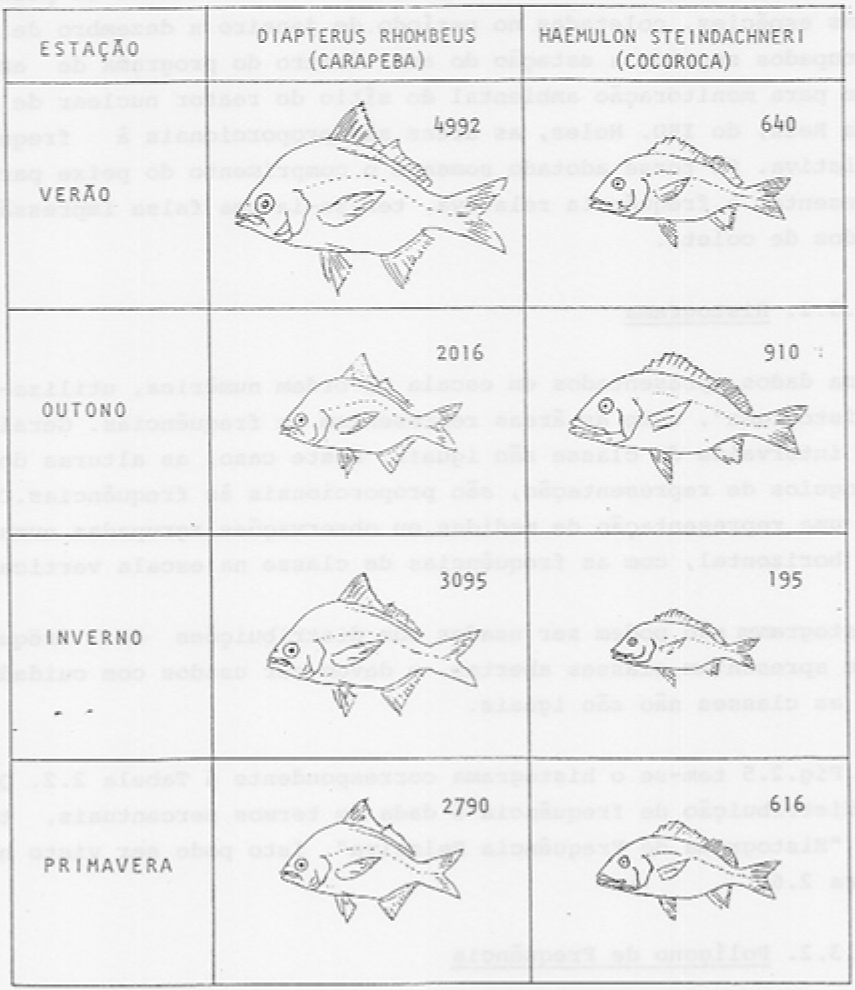
Histogramas não podem ser usados com distribuições de frequência que apresentam classes abertas, e devem ser usados com cuidado quando as classes não são iguais.

Na Fig.2.5 tem-se o histograma correspondente à Tabela 2.2. Quando a distribuição de frequência é dada em termos percentuais, tem-se um "Histograma de Frequência Relativa". Isto pode ser visto na figura 2.6.

II.3.2. Polígono de Frequência

O "Polígono de Frequência" ou "Gráfico de Linha" tem seus pontos com ordenadas iguais à frequência de classe e abscissas iguais aos símbolos das classes (pontos médios). Em geral, acrescenta-se uma classe, com frequência zero, no início e no final, a fim de dar ao gráfico uma aparência "completa".

Na Fig.2.7 tem-se o polígono de frequência correspondente aos dados da Tabela 2.2.



ESTAÇÃO	DIAPTERUS RHOMBEUS (CARAPEBA)	HAEMULON STEINDACHNERI (COCOROCA)
VERÃO	4992	640
OUTONO	2016	910
INVERNO	3095	195
PRIMAVERA	2790	616

FIG.2.4

Número de peixes coletados no Saco de Piraquara de Fora, Angra dos Reis, no período de Janeiro a Dezembro de 1982 dentro do programa de amostragem para monitoração ambiental do sítio de Angra, pelo IRD.

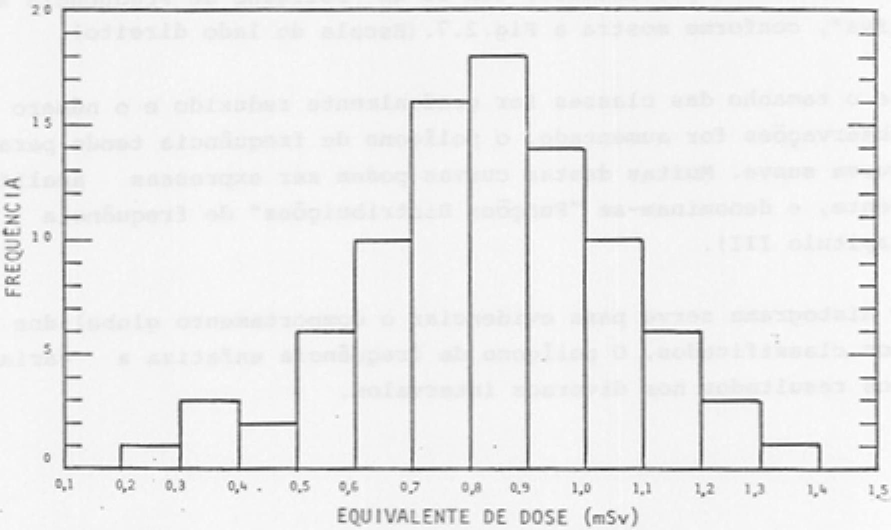

FREQUÊNCIA
0
5
10
15
20
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0
1,1
1,2
1,3
1,4
1,5
EQUIVALENTE DE DOSE (mSv)

Fig.2.5

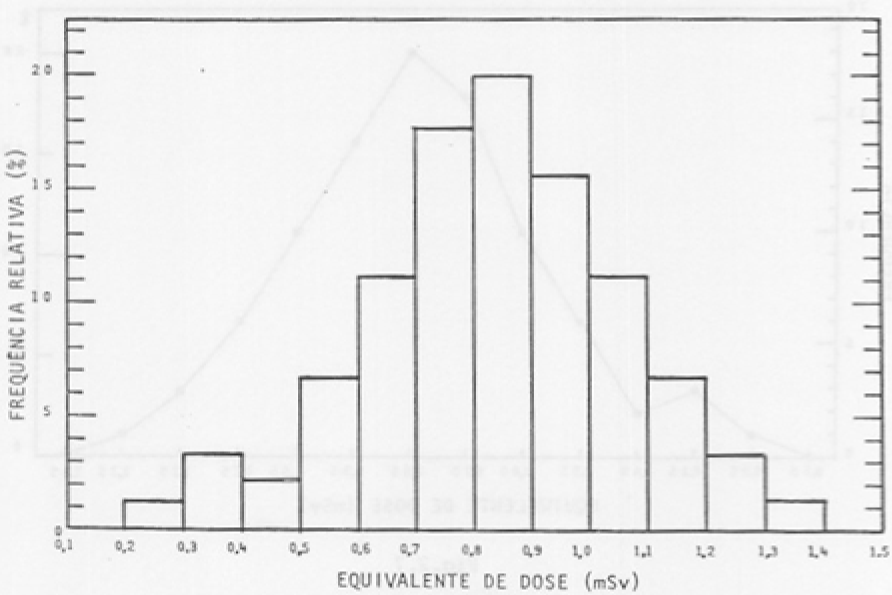

FREQUÊNCIA RELATIVA (%)
0
5
10
15
20
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0
1,1
1,2
1,3
1,4
1,5
EQUIVALENTE DE DOSE (mSv)

Fig.2.6

Da mesma forma que no histograma, se as frequências forem expressas em termos percentuais, tem-se um "Polígono de Frequência Relativa", conforme mostra a Fig.2.7.(Escala do lado direito).

Se o tamanho das classes for gradualmente reduzido e o número de observações for aumentado, o polígono de frequência tende para uma curva suave. Muitas destas curvas podem ser expressas analiticamente, e denominam-se "Funções Distribuições" de frequência (ver Capitulo III).

O histograma serve para evidenciar o comportamento global dos dados classificados. O polígono de frequência enfatiza a variação dos resultados nos diversos intervalos.

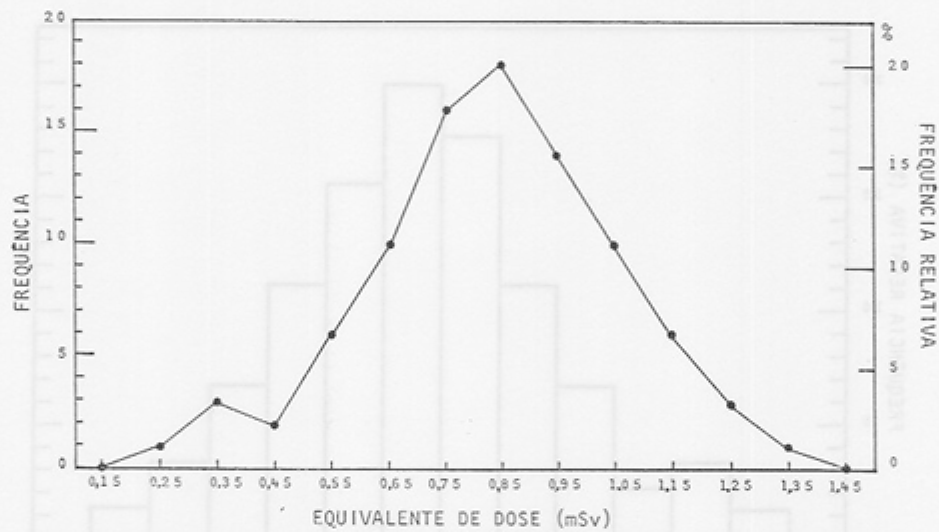


Fig.2.7

II.3.3. Ogiva

Quando se tem interesse no número de casos que se situam abaixo ou acima de um valor específico, ao invés de nos intervalos, é conveniente usar a "Distribuição de Frequência Acumulada" no lugar da usual distribuição de frequência. A frequência acumulada ou cumulativa é obtida, somando-se todas as frequências menores ou igual que o valor listado.

Para valores normalizados a 100%, tem-se uma "Distribuição de Frequência Acumulada Relativa" ou percentual. A Tabela 2.3. apresenta os valores calculados, utilizando-se os dados da Tabela 2.2.

Tabela 2.3 - Frequência Acumulada e Frequência Acumulada Relativa

CLASSE (mSv)	FREQUÊNCIA	FREQUÊNCIA RELATIVA (%)	FREQUÊNCIA ACUMULADA	FREQ.ACUMULADA RELATIVA (%)
0,20-0,30	1	1,1	1	1,1
0,30-0,40	3	3,3	4	4,4
0,40-0,50	2	2,2	6	6,6
0,50-0,60	6	6,7	12	13,3
0,60-0,70	10	11,1	22	24,4
0,70-0,80	16	17,8	38	42,2
0,80-0,90	18	20,0	56	62,2
0,90-1,00	14	15,6	70	77,8
1,00-1,10	10	11,1	80	88,9
1,10-1,20	6	6,7	86	95,6
1,20-1,30	3	3,3	89	98,9
1,30-1,40	1	1,1	90	100,0
TOTAL	90	100,0%		

A representação gráfica, tendo como ordenada a frequência acumulada ou a frequência acumulada relativa, e como abscissa, a classe ou o símbolo de classe, denomina-se "Polígono de Frequência Acumulada" ou "Polígono de Frequência Acumulada Relativa", respectivamente. Tais gráficos são também denominados de "Ogivas".

No caso da utilização do símbolo de classe como abscissa, o valor utilizado é o limite superior real de classe. Na Fig.2.8 tem-se o polígono de frequência acumulada utilizando a classe como abscissa, com os dados da Tabela 2.3, e na Fig.2.9, o polígono de frequência acumulada relativa em função dos símbolos de classe.

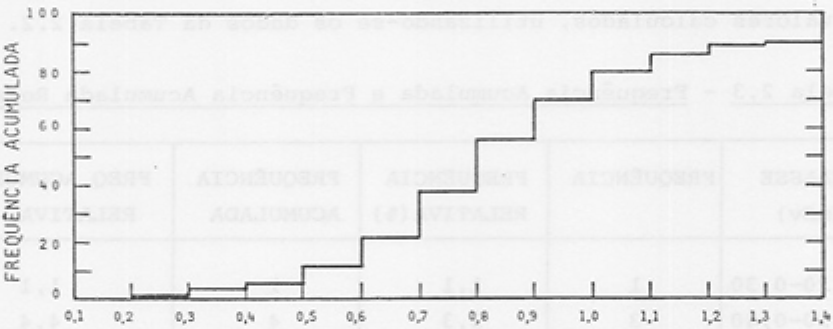


Fig.2.8

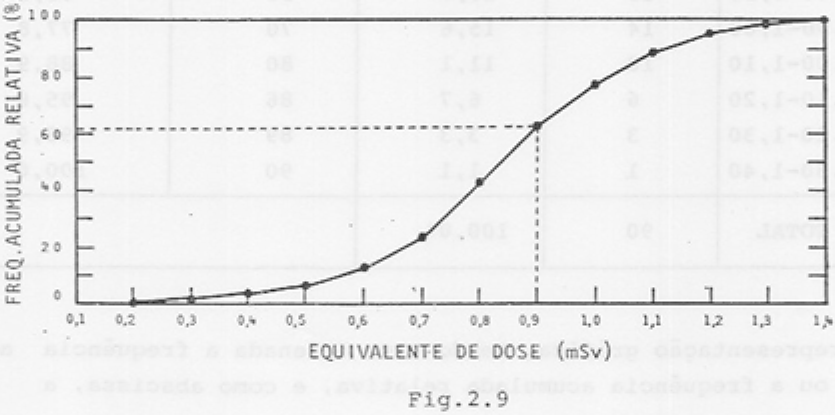


Fig.2.9

As ogivas permitem obter facilmente o percentual de medidas que tem valores menores que um valor especificado. Por exemplo,na Fig. 2.9, 62,2% dos equivalentes de dose absorvida são menores que 0,9 miliSievert.

Se se desejar os valores maiores que um valor especificado, constroi-se uma "Ogiva Reversa", obtida subtraindo-se de 100% a frequência acumulada relativa.

EXERCICIO 2.1 : CLASSIFICAÇÃO DE DADOS E REPRESENTAÇÃO GRÁFICA

Na Tabela 2.1 estão registrados 60 valores do Equivalente de Dose de um operador de gamagrafia, obtidos mensalmente por dosímetro pessoal. Realizar as seguintes operações:

1) Classificar os dados em intervalos de classe de tamanho 0,10 mSv, tendo como limite real inferior da 1a. classe, 0,40 mSv;
2) Após a Tabulação dos resultados, construir o "Histograma", representativo de Frequência versus Equivalente de Dose;
3) Construir o Histograma de Frequência Relativa x Equiv.Dose;
4) Construir o Polígono de Frequências, utilizando como simbolo de classe o ponto médio de cada classe;
5) Calcular a Frequência Acumulada e a Frequência Acumulada Relativa;
6) Construir as Ogivas ou gráficos de Frequência Acumulada x Equivalente de Dose e Frequência Acumulada Relativa x Eq.Dose, sendo que para este último, utilizar o limite real superior de classe, como símbolo de classe.

Tabela 2.1 - Valores registrados de Equivalente de Dose (mSv) de um operador de gamagrafia

0,51	0,63	0,90	0,91	0,67	0,71	0,73	0,98	1,00	0,77
0,79	0,77	0,59	0,95	0,53	0,58	0,65	0,57	0,70	0,64
0,42	0,58	0,60	0,76	0,78	0,71	0,70	0,95	0,91	1,05
0,49	0,62	0,63	0,60	0,74	0,79	0,97	0,67	0,79	1,03
0,68	0,68	0,73	0,75	0,93	1,18	0,94	0,80	0,80	0,89
0,81	0,81	0,85	0,87	0,80	0,83	0,89	0,84	0,85	0,80

EXERCÍCIO 2.2 - CLASSIFICAÇÃO DE DADOS E REPRESENTAÇÃO GRÁFICA

Os resultados da monitoração de uma mina de urânio,tipo galeria,durante um certo período de tempo,foram os seguintes:

COD.TRABALHADOR	EXPOSIÇÃO ALFA ACUMULADA (WLM)	TEMPO DE EXP. (MES)
Dinamitadores		
01	3,85	14
02	6.23	15
03	4,40	10
04	3,89	14
Engenheiros		
05	0,05	3
06	0,20	3
Eletricista		
07	1,31	12
08	0,80	8
Encanador		
09	2,26	22
10	9,16	10
11	3,69	10
Geólogo		
12	0,49	4
13	0,36	2
Guincheiro		
14	2,57	13
15	1,50	10

COD.TRABALHADOR	EXPOSIÇÃO ALFA ACUMULADA	TEMPO DE EXPOS.
Madereiro		
16	0,82	12
17	2,20	16
18	1,68	17
19	2,32	18
20	1,90	9
21	1,41	13
22	2,50	2
23	2,96	9
24	0,78	9
Mecânico		
25	3,71	16
26	0,29	9
Operador de Máquina		
27	3,36	15
28	2,83	13
29	6,33	14
30	3,92	15
Prospector		
31	1,57	8
Rolador de Sub-solo		
32	1,84	9
33	3,10	12
34	5,81	12
35	2,91	22
36	3,38	16
37	2,04	14
38	1,89	11
39	2,55	17
40	2,68	12
41	2,72	12
42	1,58	10
Supervisor de Segurança		
43	1,85	18
44	0,64	10
Proteção Radiológica		
45	1,19	11

COD.TRABALHADOR	EXPOSIÇÃO ALFA ACUMULADA (WLM)	TEMPO DE EXPOS. (MES)
Soldador		
46	0,15	10
Técnico de Mineração		
47	2,20	5
48	0,78	4
49	0,41	7
Topógrafo		
50	1,58	9

Com estes dados realizar as operações seguintes:

1. Classificar os resultados,em intervalos regulares de 0,1, para a Exposição Alfa Acumulada Média por Mes;
2. Fazer o Histograma,Exposição Alfa Acumulada Média/Mes,em unidades de (WLM/M) versus a frequencia de ocorrência;
3. Calcular e representar em gráfico a frequência acumulada em termos percentuais,e verificar se a distribuição é do tipo Normal ou Log-Normal;
4. Calcular o valor médio e o desvio padrão da média para a distribuição de dados classificados;
5. Calcular a Exposição Alfa Acumulada Média/Mes Coletiva,ou seja,em WLM/(M.PESSOA);

A Concentração de Energia Potencial Alfa no Ar(C_p): de qualquer mistura de filhos de Rn-222 e Rn-220 é a soma da energia potencial alfa de todos os átomos filhos presentes em um volume unitário de ar.

A unidade especial de medida da Concentração de energia potencial alfa é o WORKING LEVEL(WL)

$$1 \text{ WL} = 1,3 \times 10^{5} \text{ MeV/litro} = 2,08 \times 10^{-5} \text{ J/m}^3$$

1 WL corresponde,aproximadamente,à concentração de energia potencial alfa no ar de filhos de vida-média curtas de Rn-222 em equilíbrio radioativo com uma concentração de atividade de Rn-222 de 3 700 Bq/m^3.Para filhos de Rn-220 em equilibrio ,1 WL corresponde a uma concentração de 275 Bq/m^3.

III - PARÂMETROS ESTATÍSTICOS

" Have you heard about the statistics students who could not swin and drowned while crossing a stream which averaged three feet deep? "

-Anônimo-

III.1 - MÉDIA ARITMÉTICA

O modo usual de se descrever a tendência central, ou localização central, de um conjunto de dados é o valor médio, definido pela "Média Aritmética", ou simplesmente "Média". A média é útil quando se tem dados que se distribuem simetricamente em torno de um valor.

A média é obtida pela soma de todos os valores de um conjunto de observações, dividida pelo número total de valores. Assim, se x_1, x_2,...x_n representam valores de n observações, a média (média da amostra), denotada por $\bar{x}$, é dada por:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3.1)$$

onde, n é o tamanho da amostra.

Utilizando os dados da Tabela 2.1, obtem-se um equivalente de dose absorvida médio de 0,83 mSv , para o procedimento técnico referido.

A média da amostra é uma estimativa da tendência da população. Convencionalmente se utiliza,para denotar os parâmetros representativos da amostra, letras romanas, enquanto que, para a população, letras gregas.

A média μ de uma população de N ítens (observação total) é defini da do mesmo modo, ou seja,

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad (3.2)$$

Na prática, a média μ não é calculada, porque não é conveniente e nem sempre possível efetuar uma enumeração completa da população. Assim, seu valor será sempre estimado, com um certo grau de confiança, pela média da amostra.

III.2 - MÉDIA PONDERADA

Quando se trata com valores médios, é frequente levar-se em conta que nem todos os valores medidos tem a mesma importância em termos de representatividade da amostra. É necessário atribuir a eles o grau próprio de importância, associando-lhes "Pesos". Neste caso,o valor médio deverá ser calculado por uma "Média Ponderada", dada por:

$$\bar{x} = \frac{w_1x_1 + w_2x_2 + \ldots\ldots + w_nx_n}{w_1 + w_2 + \ldots. + w_n} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i} \qquad (3.3)$$

onde,

x_i = valores observados

w_i = peso associado ao valor x_i

Quando os dados estão agrupados segundo uma distribuição de frequência, a média é calculada através de (3.3), onde as frequências das classes contituem os pesos,ou seja,

$$\bar{x} = \frac{\sum_{i=1}^{m} f_ix_i}{\sum_{i=1}^{m} f_i} \qquad (3.4)$$

0,20 - 0,30	1	0,25	0,25
0,30 - 0,40	3	0,35	1,05
0,40 - 0,50	2	0,45	0,90
0,50 - 0,60	6	0,55	3,30
0,60 - 0,70	10	0,65	6,50
0,70 - 0,80	16	0,75	12,00
0,80 - 0,90	18	0,85	15,30
0,90 - 1,00	14	0,95	13,30
1,00 - 1,10	10	1,05	10,50
1,10 - 1,20	6	1,15	6,90
1,20 - 1,30	3	1,25	3,75
1,30 - 1,40	1	1,35	1,35

$\Sigma f_i = 90$ $\Sigma f_i x_i = 75,10$

portanto, $\bar{x} = 0,83$ mSv.

No caso de uma população contendo um número infinito de ítens, sua distribuição de frequência pode ser descrita analiticamente por uma função P(x), denominada "Função Distribuição". O valor médio μ é obtido através do "Valor Esperado" de x, através da expressão

$$\mu = \langle x \rangle = \frac{\int P(x)\ x\ dx}{\int P(x)\ dx} \qquad (3.5)$$

onde os limites de integração podem ir de $-\infty$ a $+\infty$, ou de 0 a $+\infty$.

Se a função distribuição for normalizada a 1, ou seja,

$$\int P(x)\ dx = 1 \qquad (3.6)$$

então,

$$\mu = \int P(x)\ x\ dx \qquad (3.7)$$

Dentre as funções distribuição, existem algumas de grande importância em estatística, como a de Poisson, de Gauss, de Student, etc que serão discutidas mais adiante.

III.3 - MEDIANA

Para dados não classificados e ordenados segundo o valor, a mediana é simplesmente o valor do ítem central. Para um número impar de observações, a mediana é diretamente determinada. Para um número par, existem 2 valores centrais. Por convenção, o valor intermediário é definido como a mediana.

Por exemplo, em 5 amostras de banana de Mambucaba, no programa pré-operacional do reator de Angra dos Reis, obteve-se os seguintes valores de atividade beta total, em pCi/g : 238, 248, 252, 262, 295. O valor mediano da atividade específica será: 252 pCi/g.

Se mais outra amostra, com resultado de 273 pCi/g for considerada, tem-se: 238, 248, 252, 262, 273, 295. A mediana, neste caso, será o valor médio de 252 e 262, ou seja, 257 pCi/g.
A mediana de dados não agrupados só serve para a localização do valor central.

Para dados agrupados, ou classificados, a mediana é obtida por interpolação, observando-se o modo de distribuição dos valores.Graficamente, a mediana é o valor da abscissa, correspondente à vertical, que divide o histograma em duas partes de mesma área. Assim, no histograma da Fig.3.1 , relativo à Tabela 3.2 para valores de posição Md = $\tilde{x}$ = 1059 mR.

A Tabela 3.2 agrupa, em classes de 200 mR, os valores de exposição em radiografias da região molar superior, registradas no cartão dosimétrico, na ponta do cone. Nestas 598 inspecções, via postal, realizadas até 1982, o valor médio da exposição de 1241 mR é 3 vezes o valor necessário para se obter uma boa radiografia num procedimento técnico correto.

Como existem 598 observações, o valor que divide o histograma em 2 partes,contendo 299 valores cada, se situa na classe 1000 - 1200mR. A mediana é determinada pela expressão:

$$Md = \tilde{x} = L_{Md} + \left(\frac{\frac{n}{2} - \Sigma f_p}{f_{Md}} \right) . j \qquad (3.8)$$

onde,

$Md = \tilde{x}$ = mediana

L_{Md} = limite inferior real da classe que contem a mediana

n = número total de observações = Σf_i

Σf_p = soma das frequências nas classes que precedem a classe que contém a mediana

j = tamanho do intervalo de classe

Tabela 3.2 - Exposição, na Ponta do Cone, em Radiologia Oral

CLASSE	FREQ.	CLASSE	FREQ.	CLASSE	FREQ.
0 - 200	12	1400 - 1600	43	2800 - 3000	10
200 - 400	56	1600 - 1800	31	3000 - 3200	7
400 - 600	57	1800 - 2000	22	3200 - 3400	10
600 - 800	80	2000 - 2200	26	3400 - 3600	4
800 - 1000	74	2200 - 2400	13	3600 - 3800	4
1000 - 1200	68	2400 - 2600	13	3800 - 4000	4
1200 - 1400	57	2600 - 2800	6	4000 - 4200	1
TOTAL: 598 inspecções					

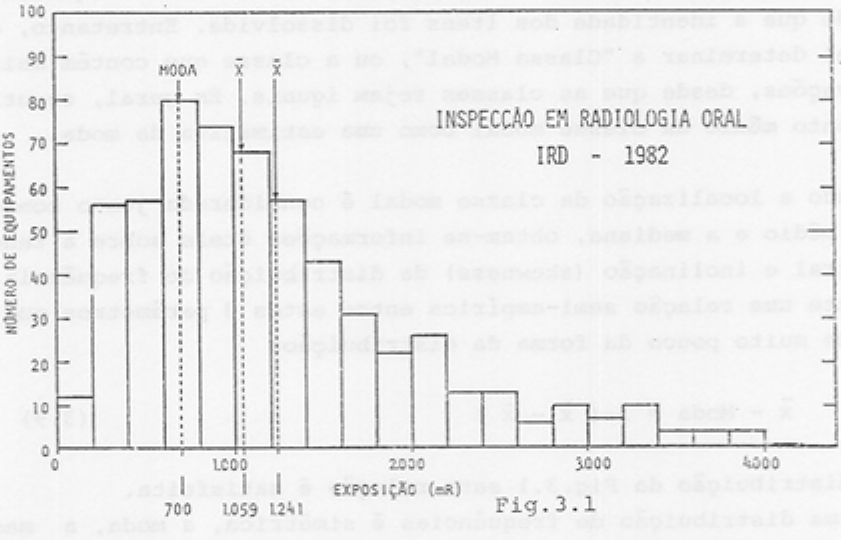


Fig.3.1

A mediana é um indicador da tendência central, livre da distorção proveniente de inclinações da distribuição de frequência. Ela é afetada pelo número de observações, mas não pelo valor individual de cada uma. Assim, num conjunto de dados em que a mediana foi determinada, se o maior valor for aumentado 100 vezes, a mediana permanece inalterada.

A desvantagem da mediana é que ela é uma média de posição e, portanto, não um conceito matemático conveniente para tratamento algébrico. Por exemplo, conhecendo-se as medianas de duas distribuições, a mediana da distribuição combinada. Para propósito de estimativa ela é mais afetada por variações na amostra.

III.4 - MODA

Outro parâmetro conceitualmente útil é a "Moda". Corresponde ao valor que apresenta a mais alta frequência. É então, o valor "mais em evidência" ou "mais em moda". Tem a vantagem de não requerer nenhum cálculo específico, somente uma contagem.

Quando os dados estão agrupados numa distribuição de frequência, não é possível especificar a observação que ocorre mais frequentemente, desde que a identidade dos ítens foi dissolvida. Entretanto, é possível determinar a "Classe Modal", ou a classe que contém mais observações, desde que as classes sejam iguais. Em geral, se utiliza o ponto médio da classe modal como uma estimativa da moda.

Quando a localização da classe modal é considerada junto com o valor médio e a mediana, obtem-se informações úteis sobre a tendência central e inclinação (skewness) da distribuição de frequência.
Existe uma relação semi-empírica entre estes 3 parâmetros que depende muito pouco da forma da distribuição:

$$\bar{x} - \text{Moda} \cong 3 \, (\bar{x} - \tilde{x}) \qquad (3.9)$$

Na distribuição da Fig.3.1 esta relação é satisfeita.
Se uma distribuição de frequências é simétrica, a moda, a mediana e a média coincidem.

Se mais que uma moda aparece, a distribuição de frequência é dita ser "Multimodal". Na Fig.3.2 se exemplifica uma distribuição bimodal.

Numa função distribuição, a moda é determinada pelo ponto de máximo da função, ou seja, onde a derivada se anula.

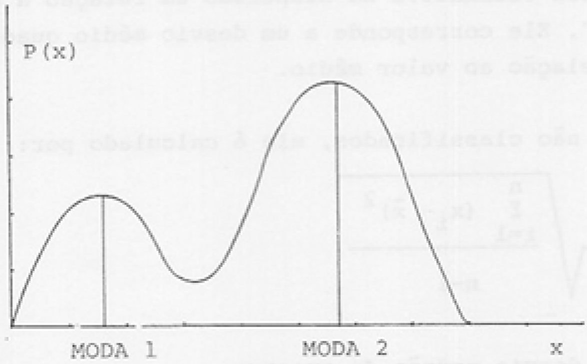


Fig.3.2

III.5 - ALCANCE

A tendência central, estimada pela média, mediana e moda, constitui importante característica descritiva dos dados estatísticos. Entretanto, varios conjuntos de dados podem apresentar médias similares, mas exibir consideravel variação nas observações individuais.

Esta variação pode ser caracterizada por parâmetros que expressam a dispersão dos valores em relação à media. A mais simples avaliação da dispersão é o "Alcance", definido como a diferença entre os valores mais alto e mais baixo, num conjunto de dados.

O alcance é fácil de calcular e de entender, mas de pouca utilidade pois, dentro dele, muitas formas alternativas de variação são possíveis. Assim, os conjuntos de dados:

Conjunto 1:	5	20	19	20	20	20	20	19	18	20	19	20	20
Conjunto 2:	5	5	5	5	15	20	15	15	15	15	20	5	15
Conjunto 3:	5	7	9	12	15	17	19	20	19	18	17	15	5

apresentam o mesmo alcance de 20 - 5 = 15, mas são bem diferentes.

III.6 - DESVIO PADRÃO

A mais eficaz estimativa da dispersão em relação à media é o "Desvio Padrão". Ele corresponde a um desvio médio quadrático de cada ponto em relação ao valor médio.

Para dados não classificados, ele é calculado por:

$$s = \sqrt{\frac{\sum_{i=1} (x_i - \bar{x})^2}{n-1}} \qquad (3.10)$$

onde,

- s = desvio padrão da amostra
- $\bar{x}$ = média da amostra
- x_i = valor de cada ítem da amostra
- n = número de ítens da amostra

A divisão por (n-1) está associado ao número de graus de liberdade. Isto porque, o cálculo do desvio padrão depende do parâmetro, que é o valor médio $\bar{x}$, extraído dos n dados.

Para dados classificados obtidos de uma amostra, o desvio padrão é obtido por,

$$s = \sqrt{\frac{\sum_{i=1}^{m} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{m} f_i}} \qquad (3.11)$$

onde,

- $\bar{x}$ = média ponderada dos valores classificados
- x_i = ponto médio das classes
- f_i = frequência de cada classe
- m = número de classes
- $\sum_{i=1}^{m} f_i$ = número total de ítens da amostra

Desenvolvendo a expressão (3.10), obtem-se uma expressão que, em algumas situações, permite o cálculo de s de um modo mais rápido,

$$s = \sqrt{\frac{n(\Sigma x^2) - (\Sigma x)^2}{n(n-1)}} \qquad (3.12)$$

O valor s^2 é denominado de "Variância" da amostra e é muito usado em testes de hipóteses como parâmetro indicativo da qualidade dos dados.

Utilizando os valores da atividade específica beta total de amostras de banana, de Mambucaba, citados em (III.3) para ilustrar o cálculo do desvio padrão, tem-se,

Tabela 3.3 - Atividade Específica Beta Total de Banana - Mambucaba

ATIVIDADE ESP.BETA TOTAL x_i	DESVIO $(x_i - \bar{x})$	DESVIO QUADRÁTICO $(x_i - \bar{x})^2$
238	-23	529
248	-13	169
252	- 9	81
262	1	1
273	12	144
295	34	1156
$\Sigma x_i = 1568$	$\Sigma(x_i - \bar{x}) = 0$	$\Sigma(x_i - \bar{x})^2 = 2080$

Portanto, usando (3.10) ou (3.12) obtem-se, para n=6,

$$s = 20 \text{ pCi/g}$$

para um valor médio de

$$\bar{x} = 261 \text{ pCi/g}$$

Assim, em termos médios, a estimativa da atividade específica beta total de bananas coletadas em Mambucaba, será:

$$\bar{x} \pm s = (261 \pm 20) \text{ pCi/g}$$

Se o desvio padrão for expresso em termos percentuais e relativo à média, tem-se o "Desvio Padrão Relativo", ou seja,

$$s(\%) = \frac{s}{\bar{x}} \cdot 100\% \qquad (3.13)$$

Entretanto, o desvio padrão s da amostra não é o único e nem o melhor indicador da dispersão dos resultados, em relação à média. Ele fornece um valor relacionado à dispersão média dos pontos. Não fornece uma avaliação direta do quanto o valor médio $\bar{x}$ pode se alterar.

O parâmetro que quantifica tal flutuação é o "Desvio Padrão da Média", denotado por $s_{\bar{x}}$ e calculado por,

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad (3.14)$$

onde n = número de ítens da amostra.

Desta maneira, o resultado da atividade específica de bananas fica melhor expresso por,

$$x \pm s_{\bar{x}} = (261 \pm 8) \text{ pCi/g}$$

É óbvio que, quando se fala em valor médio de uma grandeza é preciso adicionar algo mais do que o desvio padrão ou desvio padrão da médio. Deve-se explicitar o nível de confiança com que a estimativa, através da amostra, é fornecida. Neste contexto, o resultado da atividade específica de bananas, será dado pelo valor médio $\bar{x}$ acrescido de $\pm\, t_{\alpha/2} \cdot s_{\bar{x}}$, onde $t_{\alpha/2}$ é um fator multiplicativo que prefixa o nível de confiança (ou a margem de erro). Esta questão será discutida mais à frente.

Conforme se observa em (3.10) e (3.14), o desvio padrão s e o desvio padrão da média $s_{\bar{x}}$ possuem as mesmas unidades que o valor médio $\bar{x}$ da grandeza avaliada. Para propósitos de comparação da dispersão de conjuntos de dados de natureza diferente exige-se que eles sejam expressões em termos percentuais ou sejam adimensionais. O desvio padrão relativo s(%), dado por (3.13) possui tais requisitos e, geralmente, é denominado de "Coeficiente de Variação" da amostra, ou seja,

$$V = s(\%) = \frac{s}{\bar{x}}.100\% \qquad (3.15)$$

Quando se considera o cálculo destes parâmetros para a população, é importante se ela possui um número finito ou infinito de ítens. No primeiro caso, o "Valor Médio μ", o "Desvio Padrão σ", o"Desvio Padrão da Média σ_μ" e o "Coeficiente de Variação V " são dados respectivamente por:

$$\mu = \frac{\sum_{1}^{N} x_i}{N} \qquad (3.16)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N-1}} \qquad (3.17)$$

$$\sigma_\mu = \frac{\sigma}{\mu} \qquad (3.18)$$

$$V = \frac{\sigma}{\mu}.100\% \qquad (3.19)$$

Se N >>1, então N-1 ≈ N.

No caso de populações com número infinito de elementos, é preciso conhecer a forma analítica F(x) de distribuição das frequências, o que nem sempre é possível. Esta função, em certos casos, pode ser inferida através de ajuste de dados bem comportados, de amostras grandes, e cuidadosamente extrapolados para a população, dentro de um nível de confiança possível. Nesta perspectiva, o valor médio, a variância e o coeficiente de variação, são obtidos por,

$$\mu = \frac{\int F(x)\, x\, dx}{\int F(x)\, dx} \qquad (3.20)$$

$$\sigma^2 = \frac{\int F(x)\, (x-\mu)^2\, dx}{\int F(x)\, dx} \qquad (3.21)$$

$$V = \frac{\sigma}{\mu}\, 100\% \qquad (3.22)$$

Quando a função F(x) coincide com uma das formas de "Função Distribuição" P(x), tem-se $\int P(x)\, dx = 1$, e neste caso, μ e σ^2 são os valores esperados de x e $(x-\mu)^2$, respectivamente.

É bom salientar a diferença entre o Desvio Padrão e o Desvio Padrão da Média. O primeiro, indica, em termos médios, o desvio de cada ponto em relação à média e está associado à reprodutibilidade da forma da distribuição de frequência. Isto é, se se repetir as n medidas, deve-se encontrar a mesma distribuição, com a dispersão s.

O desvio padrão da média vincula-se à reprodutibilidade do valor médio, ou seja, se se repetir as n medidas, a forma da distribuição será igual (dentro da exatidão e precisão das medidas) e seu valor médio deve-se situar dentro do intervalo $(\bar{x} - s_{\bar{x}})$ e $(\bar{x} + s_{\bar{x}})$. Tais conceitos são ilustrados pela Fig.3.3, utilizando-se os dados da Tabela 3.1. Nesta figura, o conjunto de dados foi indicado por uma curva suave, para efeito de visualização. Como será visto mais tarde, o desvio padrão está associado à precisão da medida, enquanto que o desvio padrão da média está associado à exatidão, quando os valores medidos são referenciados a um padrão.

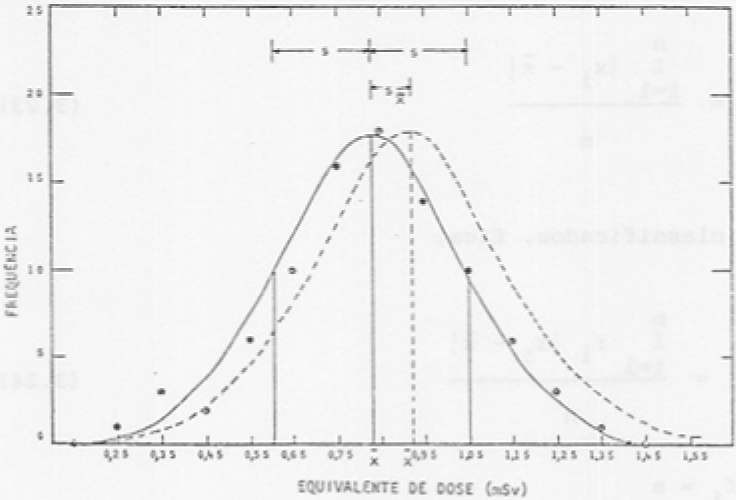


Fig.3.3

Pela Fig.3.3 podem ser inferidos o número de casos, em termos percentuais, que está incluido nos intervalos: $\bar{x} \pm s$, $\bar{x} \pm 2s$ e $\bar{x} \pm 3s$, ou seja:

$\bar{x} \pm s \quad \simeq 69\%$

$\bar{x} \pm 2s \quad \simeq 94\%$

$\bar{x} \pm 3s \quad \simeq 100\%$

onde o valor médio do equivalente de dose é $\bar{x} = 0,83$ mSv e o desvio padrão, $s = 0,22$ mSv.

Tais percentuais, contidos nestes intervalos, permitem apreciar a qualidade da amostra através do comportamento da distribuição de frequências, definindo sua randomicidade ou não.

III.7 - DESVIO MÉDIO

O "Desvio Médio" reproduz, em valor absoluto, a dispersão média dos pontos, em relação ao valor médio. Para dados não classificados e provenientes de uma amostra, tem-se,

$$\overline{\delta x} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \tag{3.23}$$

Para dados classificados, fica,

$$\overline{\delta x} = \frac{\sum_{i=1}^{m} f_i \, |x_i - \bar{x}|}{n} \tag{3.24}$$

onde, $\sum_{i=1}^{m} f_i = n$

Quando expresso em função da média e em termos percentuais, tem-se o "Desvio Médio Percentual", ou seja,

$$\overline{\delta x}(\%) = \frac{\overline{\delta x}}{\bar{x}} \; 100\% \tag{3.25}$$

Para a População, definições similares são definidas, quando possuem um número finito de elementos, mudando-se somente a notação, ou seja, $\bar{x} \rightarrow \mu$ e $n \rightarrow N$.

EXERCÍCIO 3.1 - CÁLCULO DE PARÂMETROS ESTATÍSTICOS

Um medidor de taxa de dose foi utilizado num levantamento radiométrico e os resultados, em mGy/h, de 20 medidas foram:

5, 10, 3, 4, 9, 8, 10, 6, 6, 10, 8, 7, 7, 5, 11, 6, 9, 7, 4, 8

A medida feita com um monitor padrão foi de 7,5 mGy/h, no local. Determinar:

a) a taxa de dose média;
b) a incerteza, expressa em termos de desvio padrão;
c) o desvio padrão da média;
d) o desvio percentual do resultado em relação ao padrão;
e) a incerteza total, obtida, compondo quadraticamente a incerteza do aparelho e o desvio em relação ao padrão.

EXERCÍCIO 3.2 - CÁLCULO DE PARÂMETROS ESTATÍSTICOS

Numa balança, que superestima o peso em 10%, foram feitas 10 pesagens, com os seguintes resultados, em mg:

15, 13, 17, 14, 12, 14, 16, 15, 14, 18

Qual o resultado da pesagem e qual a sua incerteza, expressa em termos de um desvio padrão?

EXERCÍCIO 3.3 - CÁLCULO DE PARÂMETROS ESTATÍSTICOS

Um operador de aparelho de raios-X diagnóstico, trabalhou 11 meses e tirou um mes de férias. Durante o ano, os resultados registrados de sua monitoração individual com filme dosimétrico, expressos em mSv, foram:

Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
15	NR	12	20	NR	40	NR	18	14	16	10	-

Determinar a dose média mensal, onde NR = dose abaixo do nível de registro

EXERCICIO 3.4 - CALCULO DE PARÂMETROS ESTATISTICOS

No programa de monitoração ambiental, pré-operacional, do sítio de Angra dos Reis, a atividade alfa-total das algas por grama de cinza, em 18 amostras, forneceu os resultados classificados na Tabela 2.4.

Definir o valor representativo do "background" da região, para a espécie de alga analisada, bem como a sua incerteza, expressos por:

a) Média aritmética e desvio padrão, supondo os valores " menores que" o limite de deteção iguais a zero;
b) Média aritmética e desvio padrão, supondo os valores " menores que" iguais ao limite de deteção;
c) Média aritmética e desvio padrão, supondo os valores "menores que" iguais a LD/2;
d) Média geométrica e desvio padrão, obtidos do gráfico de frequência acumulada relativa com distribuição log-normal;
e) Comparar, interpretar os resultados e padronizar um modo de cálculo.

Tabela 2.4 - Contagem Alfa-Total de Algas - CNAAA -1979/81

i	CONTAGEM (pCi/g)	i	CONTAGEM (pCi/g)
1	<1,0	10	3,8
2	<1,0	11	4,3
3	<2,5	12	5,7
4	<2,5	13	6,6
5	<2,6	14	7,0
6	<2,6	15	9,3
7	<2,6	16	9,5
8	3,2	17	11,7
9	3,2	18	12,4

EXERCICIO 3.5 : CÁLCULO DE PARÂMETROS ESTATÍSTICOS

Os resultados da monitoração individual de um grupo de 60 operadores de uma instalação foram classificados em intervalos de 0,10 mSv e estão expressos na Tabela 1.

Tabela 1.

CLASSE (mSv)	FREQUÊNCIA
0,40-0,50	2
0,50-0,60	6
0,60-0,70	11
0,70-0,80	15
0,80-0,90	13
0,90-1,00	9
1,00-1,10	3
1,10-1,20	1

Calcular:

a) Frequência relativa
b) Frequências Acumulada
c) Frequência Acumulada Relativa
d) Construir o Histograma de Frequência Relativa x Dose Equivalente;
e) Utilizando o papel probabilístico, fazer o gráfico da Frequencia Acumulada Relativa x Dose Equivalente;
f) Calcular a Média Ponderada, utilizando o histograma ou a Tabela;
g) O Desvio Padrão;
h) O Desvio Padrão da Média;
i) O número de casos compreendidos nos intervalos $\overline{X} \pm s$, $\overline{X} \pm 2s$ $\overline{X} \pm 3s$ utilizando o gráfico no papel probabilístico;
j) A probabilidade de um operador ser exposto à mais de 1,0 mSv por mês, utilizando o mesmo procedimento de trabalho.

IV - FUNÇÕES DISTRIBUICÃO

" The big problem with management science models is that managers practically never use them. "

- John P.S. Little -

As distribuições de frequência encontradas na vida real podem, com boa aproximação, serem simuladas por funções matemáticas. Algumas destas funções apresentam um grupo de propriedades especiais, que permitem a descrição do comportamento estatístico de muitos fenômenos, e são denominadas de "Função Distribuição".

As funções distribuições podem operar com variáveis discretas ou contínuas. Entre as discretas, as de maior interesse são : Binomial, Multinomial e de Poisson; e entre as contínuas: de Gauss (ou Normal), Log-Normal, Qui-Quadrado, de Student e Distibuição F.

IV.1 - DISTRIBUIÇÕES DISCRETAS

IV.1.1 - DISTRIBUIÇÃO BINOMIAL

Muitas amostras ou populações consistem somente de elementos de duas espécies: par-ímpar, macho-fêmea, infestado-limpo, quente- frio, morto-vivo, etc. O interesse da investigação está então na proporção ou porcentagem, ou seja, no número de ítens que estão em uma das duas classes.

Uma amostra ou população desse tipo (de 2 classes) tem uma estrutura muito simples; ela pode ser descrita, dando a proporção p do número total de ítens que estão em uma classe.

Um acontecimento de ocorrência certa, apresenta uma probabilidade p=1, enquanto um acontecimento impossível, p=0. A probabilidade varia, então, entre 0 e 1.

Se um evento pode ocorrer ou não, ele se denomina aleatório ou casual. Nesse caso, para um número N de acontecimentos aleatórios,caso não exista razão que justifique uma preferência para um deles isoladamente, a probabilidade de ocorrência será,

$$p = \frac{1}{N} \qquad (4.1)$$

Se agora, um acontecimento aleatório A aparece como consequência de qualquer r ocorrências, pertencentes ao total N de acontecimentos, a probabilidade do acontecimento A é dada por,

$$p = \frac{r}{N} \qquad (4.2)$$

Se se efetuam N testes independentes, e se em cada um o evento A apresenta uma probabilidade p de ocorrência,a probabilidade de que A apareça r vezes é igual a,

$$P(r) = C_N^r \, p^r \, (1-p)^{N-r} \, , \quad r = 0,1,\ldots N \qquad (4.3)$$

onde,

$$C_N^r = \frac{N\ (N-1)\ (N-2)\ldots\ldots(N-r+1)}{r\ (r-1)\ (r-2)\ldots\ldots(2)\ (1)} \qquad (4.4)$$

e onde p é a probabilidade de sucesso. Naturalmente, $(1-p) = q$, é a probabilidade de fracasso, de modo que, $p + q = 1$.

Uma distribuição do tipo (4.3) é chamada de "Binomial", porque as probabilidades são os termos sucessivos da expansão da expressão binomial $(p+q)^N$.

Essa distribuição dá a probabilidade de encontrar exatamente r sucessos em N tentativas, quando a probabilidade p de sucesso em cada tentativa individual é constante. Desta forma, para definir uma distribuição binomial, são necessários os parâmetros:

N inteiro e positivo

p $0 \leq p \leq 1$

Em resumo, uma distribuição binomial é obtida de uma experiência consistindo de um número inteiro de tentativas, para as quais,

a).existem exatamente só 2 saídas possíveis, para cada tentativa;

b).as probabilidades permanecem constantes de tentativa para tentativa; e

c).tentativas sucessivas são independentes.

Na Tabela 4.1 estão os valores de P(r), para N = 8, e diversos de valores de p.

Tabela 4.1 - Distribuição Binomial: N=8

r	p = 0,1	p = 0,3	p = 0,5	p = 0,7	p = 0,9
0	0,4305	0,0576	0,0039	0,0001	0,0000
1	0,3826	0,1977	0,0312	0,0012	0,0000
2	0,1489	0,2965	0,1094	0,0100	0,0000
3	0,0331	0,2541	0,2188	0,0467	0,0004
4	0,0046	0,1361	0,2734	0,1361	0,0046
5	0,0004	0,0467	0,2188	0,2541	0,0331
6	0,0000	0,0100	0,1094	0,2965	0,1489
7	0,0000	0,0012	0,0312	0,1977	0,3826
8	0,0000	0,0001	0,0039	0,0576	0,4305

Esses valores foram obtidos usando a equação (4.3). Por exemplo, para p = 0,5 e r = 3, tem-se,

$$P(r=3) = C_8^3 \ (0,5)^3 \ (1-0,5)^{8-3}$$

onde,

$$C_8^3 = \frac{8.7.6}{3.2.1} = 56$$

e então:

$$P(r=3) = 56. \ (0,5)^8 = 0,2188$$

A representação gráfica para p = 0,1, 0,3 e 0,5 está na Fig.4.1. É interessante observar que, quando N.p = inteiro, a distribuição apresenta a média igual à moda.

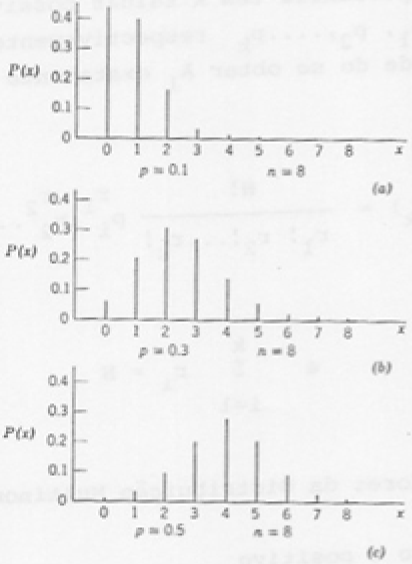


Fig.4.1

Se uma experiência, que obedece a distribuição binomial, é repetida N vezes, e a probabilidade de sucesso em uma tentativa é p,a média da distribuição será,

$$\mu = N.p \tag{4.5}$$

a variância,

$$\sigma^2 = N\,p\,(1-p) \tag{4.6}$$

e o desvio padrão,

$$\sigma = \sqrt{N\,p\,(1-p)} \tag{4.7}$$

IV.1.2 - DISTRIBUIÇÃO MULTINOMIAL

A generalização da distribuição binomial, para o caso em que mais de 2 saídas são possíveis em uma experiência, é a "Distribuição Multinomial". Se uma experiência tem k saídas possíveis: $A_1, A_2 \ldots A_k$, com probabilidades $p_1, p_2, \ldots p_k$ respectivamente, e é repetida N vezes, a probabilidade de se obter A_i exatamente r_i vezes, é dada por:

$$P(r_1, r_2, \ldots, r_k) = \frac{N!}{r_1!\, r_2! \ldots r_k!}\, p_1^{r_1} . p_2^{r_2} \ldots p_k^{r_k} \qquad (4.8)$$

onde,

$$\sum_{i=1}^{k} p_i = 1 \qquad \text{e} \qquad \sum_{i=1}^{k} r_i = N$$

Os parâmetros definidores da Distribuição Multinomial são:

N inteiro e positivo

k inteiro e positivo

$p_1 \geq 0,\ p_2 \geq 0, \ldots\ p_k \geq 0$ com $\sum_{i=1}^{k} p_i = 1$

A média de uma distribuição multinomial é dada por,

$$\mu(r_i) = N\, p_i \qquad (4.9)$$

a variância, por,

$$\sigma^2(r_i) = N\, p_i\, (1-p_i) \qquad (4.10)$$

e o desvio padrão,

$$\sigma(r_i) = \sqrt{N\, p_i\, (1-p_i)} \qquad (4.11)$$

IV.1.3 - DISTRIBUIÇÃO DE POISSON

A Distribuição de Poisson pode ser obtida a partir da distribuição binomial (eq.4.3), fazendo $N \to \infty$, e $p \to 0$. Ela constitui uma boa aproximação da distribuição binomial quando, por exemplo, p é menor que 0,10 e N é pelo menos 1000 p, ou quando (1-p) é menor que 0,10, de forma que o valor médio N.p permaneça constante e igual à μ.

A necessidade de tentativas sucessivas serem independentes permanece.

Colocando:

$$p = \frac{\mu}{N} \quad e \quad q = 1-p = 1 - \frac{\mu}{N} \quad \text{em (4.3) tem-se,}$$

$$P(r) = \frac{N\,(N-1)\ldots\,(N-r+1)}{r!}\,(1-\tfrac{\mu}{N})^{N-r}\,.\,(\tfrac{\mu}{N})^{r} \qquad (4.12)$$

O lado direito da equação pode ser rearrumado como,

$$(1-\tfrac{\mu}{N})^{N}\,(\tfrac{\mu^{r}}{r!})\,(1-\tfrac{1}{N})\,(1-\tfrac{2}{N})\ldots(1-\tfrac{r-1}{N})\,(1-\tfrac{\mu}{N})^{-r}$$

Para $N \to \infty$, $(1-\frac{\mu}{N})^{N} \to e^{-\mu}$; cada um dos (r-1) termos (1-1/N),(1-2/N),...,(1-(r-1)/N) tende a 1, assim como $(1-\mu/N)^{-r}$. Portanto, o limite de P(r), quando $N \to \infty$ será:

$$P(r) = \frac{\mu^{r}.\,e^{-\mu}}{r!} \qquad (4.13)$$

A Distribuição de Poisson é definida como (4.13) onde μ(número real e positivo) é o parâmetro definidor. Note que,

$$\sum_{r=0}^{\infty} P(r) = 1 \qquad (4.14)$$

O fato de r poder tomar um conjunto infinito de valores: 0,1,2,3.. não é problema, desde que as probabilidades tornam-se desprezíveis após relativamente poucos valores de r.

A Distribuição de Poisson é especialmente aplicada em problemas de física, como por exemplo, no registro de contagens por um detetor por unidade de tempo.

Na Fig.4.2 estão apresentados 3 exemplos de distribuição de Poisson para = 0,5 , 1,0 e 6,0.

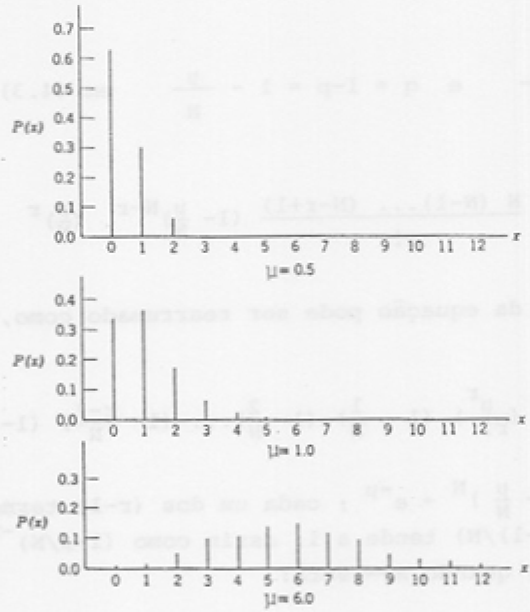


Fig.4.2

Quando o valor médio μ aumenta, observa-se uma tendência à simetria.

Para se calcular P(r) usando,(4.13), com μ = 6,0 e r = 6, tem-se,

$$P(r=6) = \frac{6^6 . e^{-6}}{6.5.4.3.2.1} = 0{,}161$$

A probabilidade para r> 6 é obtida da mesma forma, ou seja,

$$P(r\ 6) = \sum_{r=6}^{\infty} P(r) = \sum_{r=6}^{\infty} \frac{e^{-6}\,6^6}{6!} =$$

$$= 1 - \sum_{r=0}^{5} \frac{e^{-6}6^r}{r!} = 1 - 0{,}0026 - 0{,}0149 - 0{,}0446 -$$

$$- 0{,}0892 - 0{,}1339 - 0{,}1606 = 0{,}5543$$

Na Distribuição de Poisson,

$$\begin{aligned} \text{Média} &= \mu \\ \sigma^2 &= \mu \\ \sigma &= \sqrt{\mu} \end{aligned} \qquad (4.15)$$

IV.1.4 - TEOREMA DE CHEBYSHEV

Intuitivamente , a variância e o desvio padrão de uma distribuição de probabilidade medem o tamanho esperado das flutuações da correspondente variável aleatória, em relação à média. Quando o desvio padrão é grande, tem-se valores mais afastados da média, indicando como será visto, uma menor precisão das medidas. Esta importante idéia é expressa formalmente pelo Teorema de Chebyshev, que diz:

> "A probabilidade que uma variável randômica tome um valor dentro de k desvios padrões da média é, pelo menos, $(1- \frac{1}{k^2})$"

Assim, a probabilidade de se obter um valor dentro 2 desvios padrões em relação à média (pertencente ao intervalo $\mu- 2\sigma$ e $\mu+2\sigma$)

será de pelo menos, $(1-1/2^2) = 0,75$; da mesma forma para k=3, tem-se, $(1-1/3^2) = 0,89$. Nas distribuições de variáveis randômicas,normalmente os valores compreendidos dentro destes intervalos são maiores que estes, porém, tais valores constituem limites inferiores dos percentuais contidos neles.

IV.1.5 - COMPARAÇÃO ENTRE DISTRIBUIÇÕES DISCRETAS

Observando essas 3 distribuições, se enfrenta o problema de quando escolher uma delas. Essa escolha é, naturalmente, básica, e pode ser feita com o auxílio da Tabela 4.2.

Tabela 4.2 - Comparação entre Distribuições Discretas

DISTRIBUIÇÃO	PARÂMETROS	VARIÁVEIS	p	N
Binomial	N, inteiro positivo p, $0 \leq p \leq 1$	r, inteiro positivo $<N$	fixo	especificado
Multinomial	N, inteiro positivo k, inteiro positivo $p_1 \geq 0, p_2 \geq 0, \ldots p_k \geq 0$ com $\Sigma p_i = 1$	r_i, i=1,2,...k inteiro positivo $<N$	fixo	especificado
Poisson	μ, real e positivo	r, inteiro positivo	$p \to 0$	desconhecido mas muito grande

IV.2 - DISTRIBUIÇÕES CONTÍNUAS

IV.2.1 - DISTRIBUIÇÃO DE GAUSS OU NORMAL

As variáveis estudadas em IV.1 eram discretas, ou seja, podiam ser contadas. As variáveis contínuas podem ser subdivididas infinita mente, tais como o peso, temperatura, tempo, embora possam ser consideradas discretas pela medida por uma unidade padrão.

A distribuição de Gauss é a mais importante distribuição para variável contínua, pois está associada à randomicidade dos eventos. Muitos dados observados podem ser descritos por uma distribuição deste tipo, embora, no estrito senso, nenhuma distribuição de dados experimentais se estende de $-\infty$ a $+\infty$, somente a população. Algumas vezes, a suposição de distribuição de Gauss é tomada como base para análise dos dados e para o desenvolvimento de estudos estatísticos.

Quando a distribuição de Gauss é normalizada a 1 ela possui o nome de "Distribuição Normal". Nela, cada valor da variável contínua corresponde uma "densidade de probabilidade". A expressão analítica da distribuição normal é,

$$P(X) = N(\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(X-\mu)^2}{2\sigma^2}} \qquad (4.16)$$

onde, $-\infty < X < +\infty$.

O gráfico dessa função é dado pela Fig.4.3, onde se observa o valor de $N(\mu,\sigma^2)$ se aproximar de zero quando X se aproxima de $-\infty$, ou de $+\infty$, e sua simetria em relação à $X = \mu$.

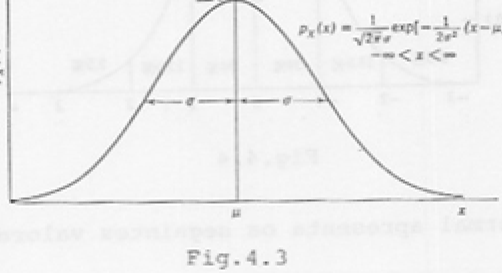


Fig.4.3

Para $X = \mu \pm \sigma$ a curva apresenta 2 pontos de inflexão.

A distribuição dada em (4.16) representa uma família de curvas dependentes de 2 parâmetros: μ(real) e σ(real positivo), que são dados na mesma unidade particular. Para evitar certas dificuldades matemáticas, é comum se fazer uso da "Distribuição Normal Padrão", em que $\mu = 0$ e $\sigma = 1$. Isto é feito através de uma mudança de variáveis, onde o valor de X, na nova distribuição, será representado pelo seu desvio em relação à média, medido em unidades de desvio padrão, ou seja,

$$u = \frac{X - \mu}{\sigma} \qquad (4.17)$$

Com a distribuição centrada em $u = 0$ e com desvio padrão $\sigma = 1$, tem-se,

$$P(u) = N(o,1) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{u^2}{2}} \;;\quad -\infty < u < \infty \qquad (4.18)$$

Assim, qualquer distribuição normal padrão tem a forma apresentada na Fig.4.4, com as percentagens das áreas contidas nos intervalos e a curva.

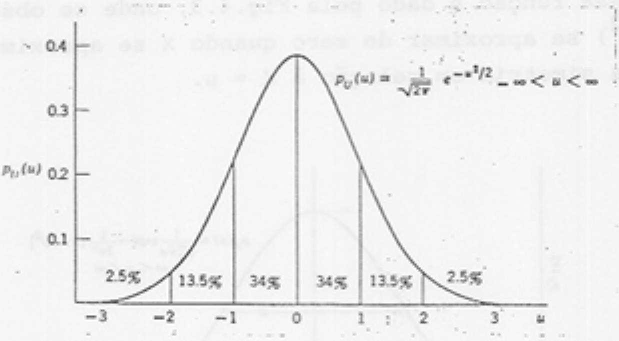


Fig.4.4

A distribuição normal apresenta os seguintes valores para seus parâ metros característicos:

Média = μ

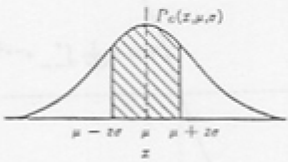
Variância = σ^2

Desvio Padrão = σ

Relembrando o teorema de Chebyshev em (IV.1.4), a distribuição normal apresenta os seguintes percentuais de área sob a curva:

INTERVALO	PERCENTUAL
$(u-\sigma) - (u+\sigma)$	68,269
$(u-2\sigma) - (u+2\sigma)$	95,449
$(u-3\sigma) - (u+3\sigma)$	99,730

<u>Tabela 4.3 - Valores dos Percentuais contidos nos Intervalos $u \pm z\sigma$</u>



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0	.00798	.01596	.02393	.03191	.03988	.04784	.05581	.06376	.07171
0.1	.07966	.08759	.09552	.10343	.11134	.11924	.12712	.13499	.14285	.15069
0.2	.15852	.16633	.17413	.18191	.18967	.19741	.20514	.21284	.22052	.22818
0.3	.23582	.24344	.25103	.25860	.26614	.27366	.28115	.28862	.29605	.30346
0.4	.31084	.31819	.32551	.33280	.34006	.34729	.35448	.36164	.36877	.37587
0.5	.38292	.38995	.39694	.40389	.41080	.41768	.42452	.43132	.43809	.44481
0.6	.45149	.45814	.46474	.47131	.47783	.48431	.49075	.49714	.50350	.50981
0.7	.51607	.52230	.52847	.53461	.54070	.54674	.55274	.55870	.56461	.57047
0.8	.57629	.58206	.58778	.59346	.59909	.60467	.61021	.61570	.62114	.62653
0.9	.63188	.63718	.64243	.64763	.65278	.65789	.66294	.66795	.67291	.67783
1.0	.68269	.68750	.69227	.69699	.70166	.70628	.71085	.71538	.71985	.72428
1.1	.72866	.73300	.73728	.74152	.74571	.74985	.75395	.75799	.76199	.76595
1.2	.76985	.77371	.77753	.78130	.78502	.78869	.79232	.79591	.79945	.80294
1.3	.80639	.80980	.81316	.81647	.81975	.82298	.82616	.82930	.83240	.83546
1.4	.83848	.84145	.84438	.84727	.85012	.85293	.85570	.85843	.86112	.86377
1.5	.86638	.86895	.87148	.87397	.87643	.87885	.88123	.88358	.88589	.88816
1.6	.89039	.89259	.89476	.89689	.89899	.90105	.90308	.90507	.90703	.90896
1.7	.91086	.91272	.91456	.91636	.91813	.91987	.92158	.92326	.92491	.92654
1.8	.92813	.92969	.93123	.93274	.93422	.93568	.93711	.93851	.93988	.94123
1.9	.94256	.94386	.94513	.94638	.94761	.94882	.95000	.95116	.95229	.95340
2.0	.95449	.95556	.95661	.95764	.95864	.95963	.96059	.96154	.96247	.96338
2.1	.96426	.96513	.96599	.96682	.96764	.96844	.96922	.96999	.97074	.97147
2.2	.97219	.97289	.97358	.97425	.97490	.97555	.97617	.97679	.97739	.97797
2.3	.97855	.97911	.97965	.98019	.98071	.98122	.98172	.98221	.98268	.98315
2.4	.98360	.98404	.98448	.98490	.98531	.98571	.98610	.98648	.98686	.98722
2.5	.98758	.98792	.98826	.98859	.98891	.98922	.98953	.98983	.99012	.99040
2.6	.99067	.99094	.99120	.99146	.99171	.99195	.99219	.99241	.99264	.99285
2.7	.99306	.99327	.99347	.99366	.99385	.99404	.99422	.99439	.99456	.99473
2.8	.99489	.99504	.99520	.99534	.99549	.99563	.99576	.99589	.99602	.99615
2.9	.99627	.99638	.99650	.99661	.99672	.99682	.99692	.99702	.99712	.99721

	.00	.10	.20	.30	.40
3.0	.9973002	.9980648	.9986257	.99903315	.99932614
3.5	.99953474	.99968178	.99978440	.99985530	.999903805
4.0	.999936656	.999958684	.999973308	.999982920	.999989174
4.5	.9999932043	.9999957748	.9999973982	.9999984132	.99999904149
5.0	.99999942657	.99999966024	.99999980061	.99999988410	.99999993327
5.5	.99999996193	.99999997847	.99999998793	.99999999328	.99999999627

Uma distribuição muito utilizada é a " Distribuição Normal Padrão Cumulativa", que fornece o valor do percentual de eventos sob a curva, de $-\infty$ até o valor de X em questão. Analiticamente isto é expresso por,

$$\Phi(u) = \int_{-\infty}^{\infty} N(0,1)\, du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2}\, du \qquad (4.19)$$

A representação gráfica de (4.19) é apresentada na Fig.4.5.

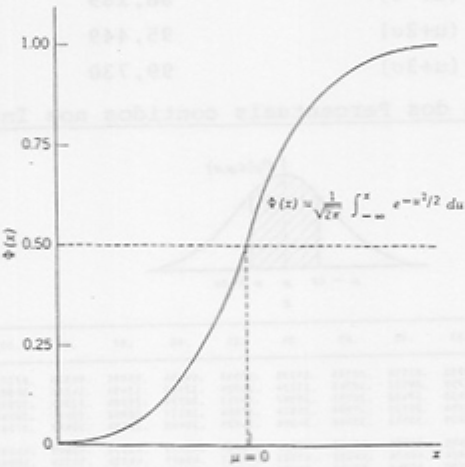


Fig.4.5

Este gráfico feito num "Papel Probabilístico" corresponde à uma reta. Desta forma, esta representação é bastante útil para verificar se os dados experimentais se distribuem como uma distribuição de Gauss, ou seja, se isso ocorrer o gráfico de frequências acumuladas percentuais, constitui uma reta. Na Fig.4.6 tem-se uma representação deste tipo, e na Fig.4.7 a representação gráfica dos dados correspondentes à Tabela 2.3 e Fig.2.9. É bom relembrar que as frequências acumuladas, nesta representação, correspondem aos limites superiores reais de cada classe.