INE.EAD

INSTITUTO NACIONAL DE ENSINO

PÓS-GRADUAÇÃO

Lato Sensu

ESTATÍSTICA APLICADA

SUMÁRIO

Introdução à Estatística 5
Importância da Estatística 6
Grandes áreas da Estatística 7
Amostragem 8
População e amostra 8
Estatística Descritiva 9
Estatística Inferencial (ou Indutiva) 10
Probabilidade 10
Fases do Método Estatístico 10
Definição do problema 11
Planejamento 11
Coleta dos dados 12
Apuração dos dados 12
Apresentação dos dados 13
Análise e interpretação dos dados 13
Séries Estatísticas 14
Frequências 39
Medidas de Posição 43
Probabilidade 61
Experimento aleatório, espaço amostral e eventos 61
Espaço Amostral 62
Probabilidade: Definição Clássica; Probabilidade e frequência relativa 65
Tipos de eventos 66
Axiomas de Probabilidade 67
Teoria da Contagem 68

Probabilidade Condicional e Independência de Eventos 69
Teorema do Produto 70
Independência Estatística 70
Regra de Bayes 71
Variáveis Aleatórias 73
Conceito de variável aleatória 74
Distribuição de probabilidade 75
Função de densidade de probabilidade 75
Esperança matemática, variância e desvio padrão: propriedades 75
Distribuições discretas: Bernoulli, Binomial e Poisson 76
Distribuições Discretas de Probabilidade 76
Distribuição de Bernoulli 77
Principais características 77
Distribuição Binomial 77
Distribuição de Poisson 79
Distribuição Normal 83
Propriedades da distribuição normal 84
A distribuição normal padrão 86
Inferência Estatística 89
População e amostra; Estatísticas e parâmetros; 89
Distribuições amostrais 89
Distribuição Amostral 90
Distribuição amostral da Média 90
Estimação 92
Estimação pontual 92
Testes de Hipóteses 96
Principais conceitos 96
Testes de Hipóteses para a Média de Populações normais com variâncias (σ^2) conhecidas 98

Testes Unilateral (Monocaudal) à Esquerda 100
Testes Unilateral (Monocaudal) à Direita 100
Correlação e Regressão linear 102
Relações Funcionais 102
Relações Estatísticas e Correlações 102
Diagrama de dispersão 103
Correlação Linear 103
Coeficiente de Correlação Linear (r) 104
Regressão – Reta de regressão (ou reta de mínimos quadrados ou reta de ajuste) 108
Referências 112

Introdução à Estatística

A palavra estatística lembra, à maioria das pessoas, recenseamento. Os censos existem há milhares de anos e constituem um esforço imenso e caro feito pelos governos, com o objetivo de conhecer seus habitantes, sua condição socioeconômica, sua cultura, religião, etc. Portanto, associar estatística a censo é perfeitamente correto do ponto de vista histórico, sendo interessante salientar que as palavras estatística e estado têm a mesma origem latina: status.

A estatística é também comumente associada às pesquisas de opinião pública, aos vários índices governamentais, aos gráficos e às médias publicados diariamente na imprensa. Na realidade, entretanto, a estatística engloba muitos outros aspectos, sendo fundamental na análise de dados provenientes de quaisquer processos onde exista variabilidade.

É possível distinguir duas concepções para a palavra ESTATÍSTICA: no plural (estatísticas), indica qualquer coleção de dados numéricos, reunidos com a finalidade de fornecer informações acerca de uma atividade qualquer. Assim, por exemplo, as estatísticas demográficas referem-se aos dados numéricos sobre nascimentos, falecimentos, matrimônios, desquites, etc. As estatísticas econômicas consistem em dados numéricos relacionados com emprego, produção, vendas e com outras atividades ligadas aos vários setores da vida econômica. No singular (Estatística), indica a atividade humana especializada ou um corpo de técnicas, ou ainda uma metodologia desenvolvida para a coleta, a classificação, a apresentação, a análise e a interpretação de dados quantitativos e a utilização desses dados para a tomada de decisões.

Importância da Estatística

O mundo está repleto de problemas. Para resolvermos a maioria deles, necessitamos de informações. Mas, que tipo de informação? Que quantidade de informações? Após obtê-las, que fazer com elas? A Estatística trabalha com essas informações, associando os dados ao problema, descobrindo como e o que coletar, assim capacitando o pesquisador (ou profissional ou cientista) a obter conclusões a partir dessas informações, de tal forma que possam ser entendidas por outras pessoas.

Portanto, os métodos estatísticos auxiliam o cientista social, o economista, o engenheiro, o agrônomo e muitos outros profissionais a realizarem o seu trabalho com mais eficiência.

A Estatística é uma parte da Matemática que fornece métodos para a coleta, organização, descrição, análise e interpretação de dados, viabilizando a utilização dos mesmos na tomada de decisões.

Vejamos alguns exemplos:

- Os estatísticos do governo conduzem censos de população, moradia, produtos industriais, agricultura e outros. São feitas compilações sobre vendas, produção, inventário, folha de pagamento e outros dados das indústrias e empresas. Essas estatísticas informam ao administrador como a sua empresa está crescendo, seu crescimento em relação a outras empresas e fornece-lhe condições de planejar ações futuras. A análise dos dados é muito importante para se fazer um planejamento adequado.
- Na era da energia nuclear, os estudos estatísticos têm avançado rapidamente e, com seus processos e técnicas, têm contribuído para a organização de empresas e utilização dos recursos do mundo moderno.

Em geral, as pessoas, quando se referem ao termo estatística, desconhecem que o aspecto essencial é o de proporcionar métodos inferenciais, que permitam conclusões que transcendam os dados obtidos inicialmente.

Grandes áreas da Estatística

Para fins de apresentação, é usual se dividir a estatística em três grandes áreas, embora não se trate de ramos isolados:

- Estatística Descritiva e Amostragem – Conjunto de técnicas que objetivam coletar, organizar, apresentar, analisar e sintetizar os dados numéricos de uma população, ou amostra;
- Estatística Inferencial – Processo de se obter informações sobre uma população a partir de resultados observados na amostra;
- Probabilidade - Modelos matemáticos que explicam os fenômenos estudados pela Estatística em condições normais de experimentação.
- Em estatística, utilizamos extensamente os termos: população, amostra, censo, parâmetros, estatística, dados discretos, dados contínuos, dados quantitativos e dados qualitativos; que estaremos definindo abaixo para maior compreensão:
- População: é uma coleção completa de todos os elementos a serem estudados.
- Amostra: é uma subcoleção de elementos extraídos de uma população.
- Censo: é uma coleção de dados relativos a todos os elementos de uma população.
- Parâmetros: é uma medida numérica que descreve uma característica de uma população.
- Estatística: é uma medida numérica que descreve uma característica de uma amostra.
- Dados contínuos: resultam de um número infinito de valores possíveis que podem ser associados a pontos em uma escala contínua de tal maneira que não haja lacunas.
- Dados discretos: resultam de um conjunto finito de valores possíveis, ou de um conjunto enumerável de valores.

- Dados quantitativos: consistem em números que representam contagens ou medidas.
- Dados qualitativos: podem ser separados em diferentes categorias que se distinguem por alguma característica não-numérica.

Amostragem

É o processo de escolha da amostra. É a parte inicial de qualquer estudo estatístico. Consiste na escolha criteriosa dos elementos a serem submetidos ao estudo. Geralmente, as pesquisas são realizadas através de estudo dos elementos que compõem uma amostra, extraída da população que se pretende analisar.

Exemplo: Pesquisas sobre tendências de votação

Em épocas de eleição, é comum a realização de pesquisas com o objetivo de se conhecer as tendências do eleitorado. Para que os resultados sejam de fato representativos, toma-se o cuidado de se entrevistar um conjunto de pessoas com características socioeconômicas, culturais, religiosas, etc. tão próximas quanto possível da população à qual os resultados da pesquisa serão estendidos. A escolha da amostra, a redação do questionário, a entrevista, a codificação dos dados e a apuração dos resultados são as etapas deste tipo de pesquisa.

População e amostra

O estudo de qualquer fenômeno, seja ele natural, social, econômico ou biológico, exige a coleta e a análise de dados estatísticos. A coleta de dados é, pois, a fase inicial de qualquer pesquisa.

É sobre os dados da amostra que se desenvolvem os estudos, visando a fazer inferências sobre a população.

Exemplo: Avaliação de um programa de ensino

Toma-se certo número de pares de turmas: a um conjunto de turmas ensina-se um assunto por um novo método e, ao outro, pelo método clássico. Aplica-se uma prova a ambos os grupos. As notas observadas nesses conjuntos de turmas constituem a nossa amostra. Se os resultados do novo método forem melhores, iremos aplicá-lo a todas as turmas, isto é, à população. A partir da amostra, estabelecemos o que é conveniente para a população, ou seja, fazemos uma inferência sobre a população.

Exemplo: Renda média per capita em diversas regiões do país

Toma-se um conjunto de indivíduos em cada região, escolhidos ao acaso, e sobre esse grupo são feitos os estudos. Os indivíduos assim escolhidos constituem a amostra e os resultados nela observados serão estendidos à população.

Estatística Descritiva

É a parte mais conhecida. Quem vê o noticiário, na televisão ou nos jornais, sabe quão frequente é o uso de médias, índices e gráficos nas notícias.

Exemplo: INPC (Índice Nacional de Preços ao Consumidor)

Sua construção envolve a sintetização, em um único número, dos aumentos dos produtos de uma cesta básica.

Exemplo: Anuário Estatístico Brasileiro

O IBGE publica esse anuário apresentando, em várias tabelas, os mais diversos dados sobre o Brasil: educação, saúde, transporte, economia, cultura, etc. Embora simples, fáceis de serem entendidas, as tabelas são o produto de um processo demorado e extremamente dispendioso de coleta e apuração de dados.

Exemplo: Anuário Estatístico da Embratur

A Embratur publica esse anuário apresentando, em várias tabelas e gráficos, os mais diversos dados sobre Turismo Interno e dados sobre entrada de turistas estrangeiros no Brasil.

Estatística Inferencial (ou Indutiva)

A tomada de decisões sobre a população, com base em estudos feitos sobre os dados da amostra, constitui o problema central da inferência estatística.

Exemplo: Suponha que a distribuição das alturas de todos os habitantes de um país possa ser representada por uma distribuição normal. Mas não conhecemos de antemão a média da distribuição. Devemos, pois, estimá-la.

Exemplo: Análise financeira. Os analistas financeiros estudam dados sobre a situação da economia, visando explicar tendências dos níveis de produção e de consumo, projetando-os para o futuro.

Exemplo: Ocorrência de terremotos. Os geólogos estão continuamente coletando dados sobre a ocorrência de terremotos. Gostariam de inferir quando e onde ocorrerão tremores, e qual a sua intensidade. Trata-se, sem dúvida, de uma questão complexa, que exige longa experiência geológica, além de cuidadosa aplicação de métodos estatísticos.

Probabilidade

O processo de generalização, que é característico do método indutivo, está associado a uma margem de incerteza. A existência da incerteza deve-se ao fato de que a conclusão, que se pretende obter para o conjunto de todos os indivíduos analisados quanto a determinadas características comuns, baseia-se em uma parcela do total das observações. A medida da incerteza é tratada mediante técnicas e métodos que se fundamentam na Teoria da Probabilidade. Essa teoria procura quantificar a incerteza existente em determinada situação.

Fases do Método Estatístico

Quando se pretende empreender um estudo estatístico completo, existem diversas fases do trabalho que devem ser desenvolvidas para se chegar aos resultados finais de um estudo capaz de produzir resultados válidos. As fases principais são as seguintes:

- Definição do problema
- Planejamento
- Coleta de dados
- Apuração dos dados
- Apresentação dos dados
- Análise e Interpretação dos dados

Definição do problema

A primeira fase do trabalho consiste em uma definição ou formulação correta do problema a ser estudado. Além de considerar detidamente o problema objeto do estudo, o analista deverá examinar outros levantamentos realizados no mesmo campo e que sejam análogos, uma vez que parte da informação de que se necessita pode, muitas vezes, ser encontrada nesses últimos.

Planejamento

O passo seguinte, após a definição do problema, compreende a fase do planejamento, que consiste em se determinar o procedimento necessário para se resolver o problema e, em especial, como levantar informações sobre o assunto, objeto do estudo. É preciso planejar o trabalho a ser realizado tendo em vista o objetivo que se pretende atingir. É nessa fase que será escolhido o tipo de levantamento a ser utilizado. Sob esse aspecto, pode haver dois tipos de levantamento:

- Levantamento censitário, quando a contagem for completa, abrangendo todo o universo;
- Levantamento por amostragem, quando a contagem for parcial.
- Outros elementos importantes que devem ser tratados nesta mesma fase são:
- Cronograma das atividades, através do qual são fixados os prazos para as várias fases;
- Custos envolvidos;
- Exame das informações disponíveis;

- Delineamento da amostra, etc.

Coleta dos dados

O terceiro passo é essencialmente operacional, compreendendo a coleta das informações propriamente ditas. Nesta fase do método estatístico, é conveniente estabelecer uma distinção entre duas espécies de dados:

- Dados primários – quando são publicados ou coletados pelo próprio pesquisador ou organização que os escolheu;
- Dados secundários – quando são publicados ou coletados por outra organização.

Um conjunto de dados é, pois, primário ou secundário em relação a alguém. As tabelas do Censo Demográfico são fontes primárias. Quando determinado jornal publica estatísticas extraídas de várias fontes e relacionadas com diversos setores industriais, os dados são secundários para quem desejar utilizar-se deles em alguma pesquisa que esteja desenvolvendo.

A coleta de dados pode ser realizada de duas maneiras:

- Coleta Direta – quando é obtida diretamente da fonte, como no caso da empresa que realiza uma pesquisa para saber a preferência dos consumidores pela sua marca;
- Coleta Indireta – quando é inferida a partir dos elementos conseguidos pela coleta direta, ou através do conhecimento de outros fenômenos que, de algum modo, estejam relacionados com o fenômeno em questão.

Apuração dos dados

Antes de começar a analisar os dados, é conveniente que lhes seja dado algum tratamento prévio, a fim de torná-los mais expressivos. A quarta etapa do processo é, então, a da apuração ou sumarização, que consiste em resumir os dados através de sua contagem e agrupamento. Pode ser manual, eletromecânica ou eletrônica.

Apresentação dos dados

Por mais diversa que seja a finalidade, os dados devem ser apresentados sob forma adequada, tornando mais fácil o exame do fenômeno que está sendo objeto de tratamento estatístico.

Há duas formas de apresentação ou exposição dos dados observados, que não se excluem mutuamente:

- Apresentação tabular – É uma apresentação numérica dos dados. Consiste em dispor os dados em linhas e colunas distribuídas de modo ordenado, segundo algumas regras práticas adotadas pelos diversos sistemas estatísticos. As tabelas têm a vantagem de conseguir expor, sinteticamente e em só local, os resultados sobre determinado assunto, de modo a se obter uma visão global mais rápida daquilo que se pretende analisar.
- Apresentação gráfica – É uma apresentação geométrica dos dados numéricos. Embora a apresentação tabular seja de extrema importância no sentido de facilitar a análise numérica de dados, não permite ao analista obter uma visão tão rápida, fácil e clara do fenômeno e sua variação como aquela conseguida através de um gráfico.

Análise e interpretação dos dados

Nesta última etapa, o interesse maior reside em tirar conclusões que auxiliem o pesquisador a resolver seu problema. A análise dos estatísticos está ligada essencialmente ao cálculo de medidas, cuja finalidade principal é descrever o fenômeno. Assim, o conjunto de dados a ser analisado pode ser expresso por números resumo, as estatísticas que evidenciam as características particulares desse conjunto. O significado exato de cada um dos valores obtidos através do cálculo das várias medidas estatísticas disponíveis deve ser bem interpretado. É possível mesmo, nesta fase, arriscar algumas generalizações, as quais envolverão, como mencionado anteriormente, algum grau de incerteza, porque não se pode estar seguro de que o que foi constatado para aquele conjunto de dados (a amostra) se verificará igualmente para a população.

Séries Estatísticas

Define-se série estatística como toda e qualquer coleção de dados estatísticos referidos a uma mesma ordem de classificação: quantitativa. No sentido mais amplo, série é uma sucessão de números referidos a qualquer variável. Se os números expressarem dados estatísticos, a série será chamada de série estatística.

Em sentido mais restrito, pode-se dizer que uma série estatística é uma sucessão de dados estatísticos referidos a caracteres qualitativos, ao passo que uma sucessão de dados estatísticos referidos a caracteres quantitativos configurará uma Distribuição de Frequência.

Em outros termos, a palavra série é usada normalmente para designar um conjunto de dados dispostos de acordo com um caráter variável, residindo a qualidade serial na disposição desses valores, e não em uma disposição temporal ou espacial de indivíduos.

Tabela é um quadro que resume um conjunto de observações.

Uma tabela compõe-se de:

- Corpo – conjunto de linhas e colunas que contém informações sobre a variável em estudo;
- Cabeçalho – parte superior da tabela que especifica o conteúdo das colunas;
- Coluna indicadora – parte da tabela que especifica o conteúdo das linhas;
- Linhas – retas imaginárias que facilitam a leitura, no sentido horizontal, de dados que se inscrevem nos seus cruzamentos com as colunas;
- Casa ou célula – espaço destinado a um só número;
- Título – conjunto de informações, as mais completas possíveis, respondendo às perguntas: O quê?- Quando?- Onde?- localizado no topo da tabela.
- Fonte – referência de onde se obteve os dados, colocado, de preferência, no rodapé.
- As tabelas servem para apresentar séries estatísticas. Conforme varie um dos elementos da série, podemos classificá-la em:

- Cronológicas - Tempo (fator temporal ou cronológico) – a que época se refere o fenômeno analisado;
- Geográficas - Local (fator espacial ou geográfico) – onde o fenômeno acontece;
- Específicas - Fenômeno (espécie do fato ou fator especificativo) – o que é descrito.

As séries também são divididas em:

- Séries Homógradas **-** aquelas em que a variável descrita apresenta variação discreta ou descontínua. São séries homógradas a série temporal, a série geográfica e a série específica;
- Séries Heterógradas **-** aquelas nas quais o fenômeno ou o fato apresenta graduações ou subdivisões. Embora fixo, o fenômeno varia em intensidade. A Distribuição de frequências ou seriação é uma série heterógrada.

Os dados estatísticos resultantes da coleta direta da fonte, sem outra manipulação senão a contagem ou medida, são chamados dados absolutos. Dados Relativos são o resultado de comparações por quociente (razões) que se estabelecem entre dados absolutos e têm por finalidade realçar ou facilitar as comparações entre quantidades.

Tipos de Séries Estatísticas Simples (ou de uma entrada)

As séries estatísticas diferenciam-se de acordo com a variação de um dos três elementos: tempo, local e fenômeno.

Série Cronológica

Também chamada de série temporal, série histórica, série evolutiva ou marcha, identifica-se pelo caráter variável do fator cronológico. Assim, deve-se ter:

Elemento variável: Época

Elementos Fixos: Local e Fenômeno

Exemplo:

Tabela 1.1 - Operadora WKX – Venda de bilhetes aéreos

– Mercado Interno – 1995

Meses	Vendas (em milhares de reais)
Janeiro	2300
Fevereiro	1800
Março	2200
Abril	2210
Maio	2360
Junho	2600
Julho	2690
Agosto	3050
Setembro	3500
Outubro	3440
Novembro	3100
Dezembro	2760
TOTAL ANUAL	31510

Fonte: Departamento de Análise de Mercado

Série Geográfica

Também chamada de série territorial, série espacial ou série de localização, identifica-se pelo caráter variável do fator geográfico. Assim, deve-se ter:

Elemento variável: Local

Elementos Fixos: Época e Fenômeno Exemplo:

Tabela 1.2 – Operadora WKX - Vendas por Unidade da Federação – 1995

Unidades da Federação	Vendas (em milhares de reais)
Minas Gerais	4000
Paraná	2230
Rio Grande do Sul	6470

Rio de Janeiro	8300
São Paulo	10090
Outros	420
TOTAL BRASIL	31510

Fonte: Departamento de Análise de Mercado

Série Específica

Também chamada de série categórica ou série por categoria, identifica-se pelo caráter variável de fator especificativo. Assim, deve-se ter:

Elemento variável: Fenômeno

Elementos Fixos: Local e Época

Exemplos:

Tabela 1.3.– Operadora WKX -

Venda de bilhetes aéreos por Linha – 1995

Linha do Produto	Vendas (em milhares de reais)
Linha A	6450
Linha B	9310
Linha C	15750
TODAS AS LINHAS	31510

Fonte: Departamento de Análise de Mercado

Tabela 1.4. Número de empregados das várias classes de salários no estado de São Paulo – 1998

Classes de Salários (R$)	Número de Empregados
Até 80	41 326
De 80 a 119	123 236
De 120 a 159	428 904
De 160 a 199	324 437

De 200 a 399	787 304
De 400 a 599	266 002
De 600 a 799	102 375
De 800 a 999	56 170
1000 e mais	103 788
TOTAL	2 233 542

Fonte: Serviço de Estatística da Previdência e Trabalho

(Dados alterados para melhor compreensão)

Tabelas Compostas (ou de dupla entrada)

As tabelas apresentadas anteriormente são tabelas estatísticas simples, onde apenas uma série está representada. É comum, todavia, haver necessidade de apresentar, em uma única tabela, mais do que uma série. Quando as séries aparecem conjugadas, tem-se uma tabela de dupla entrada. Em uma tabela desse tipo são criadas duas ordens de classificação: uma horizontal (linha) e uma vertical (coluna).

Exemplos:

a) Série específico-temporal

b) Série geográfico-temporal

Tabela 1.5 – População economicamente ativa por setor de atividades – Brasil

Setor	**População (1 000 Hab.)**		
	1940	1950	1960
Primário	8 968	10 255	12 163
Secundário	1 414	2 347	2 962
Terciário	3 620	4 516	7 525

Fonte: IPEA

Tabela 1.6 – População Indígena Brasileira

Unidade de Produção	**Produção**		
	1937	1938	1939

Acre	5 007	4 765	4 727
Amazonas	6 858	5 998	5 631
Pará	4 945	4 223	4 500
Mato Grosso	1 327	1 285	1 235
Outros Estados	333	539	337

Fonte: Anuário Estatístico do Brasil – IBGE - (Dados alterados para melhor compreensão)

Podem existir, se bem que mais raramente, pela dificuldade de representação, séries compostas de três ou mais entradas.

Observação:

Nem sempre uma tabela representa uma série estatística. Por vezes, os dados reunidos não revelam uniformidade, sendo meramente um aglomerado de informações gerais sobre determinado assunto, as quais, embora úteis, não apresentam a consistência necessária para se configurar uma série estatística.

Exemplo: Tabela com resumos de dados, mas que não representa uma série estatística.

Tabela 1.8 – Situação dos espetáculos cinematográficos no Brasil – 1967

Especificação	Dados Numéricos
Número de cinemas	2 488
Lotação dos cinemas	1 722 348
Sessões por dia	3 933
Filmes de longa metragem	131 330 488
Meia-entrada	89 581 234

Fonte: Anuário Estatístico do Brasil – IBGE

Apresentação de dados - Tabelas e Gráficos: Construção e Interpretação

A representação gráfica das séries estatísticas tem por finalidade representar os resultados obtidos, permitindo que se chegue a conclusões sobre a evolução do fenômeno ou sobre como se relacionam os valores da série. A escolha do gráfico mais

apropriado ficará a critério do analista. Contudo, os elementos simplicidade, clareza e veracidade devem ser considerados, quando da elaboração de um gráfico.

- Simplicidade – o gráfico deve ser destituído de detalhes de importância secundária, assim como de traços desnecessários que possam levar o observador a uma análise morosa ou sujeita a erros.
- Clareza – o gráfico deve possibilitar uma correta interpretação dos valores representativos do fenômeno em estudo.
- Veracidade – o gráfico deve expressar a verdade sobre o fenômeno em estudo.

Diretrizes para a construção de um gráfico:

O título do gráfico deve ser o mais claro e completo possível. Quando necessário, deve-se acrescentar subtítulos;

A orientação geral dos gráficos deve ser da esquerda para a direita;

As quantidades devem ser representadas por grandezas lineares;

Sempre que possível, a escala vertical há de ser escolhida de modo a aparecer a linha 0 (zero);

Só devem ser incluídas no desenho as coordenadas indispensáveis para guiar o olhar do leitor ao longo da leitura. Um tracejado muito cerrado dificulta o exame do gráfico;

A escala horizontal deve ser lida da esquerda para a direita, e a vertical de baixo para cima;

Os títulos e marcações do gráfico devem ser dispostos de maneira que sejam facilmente lidos, partindo da margem horizontal inferior ou da margem esquerda.

Leitura e interpretação de um gráfico:

Declarar qual o fenômeno ou fenômenos representados, a região considerada, o período de tempo, a fonte dos dados, etc;

Examinar o tipo de gráfico escolhido, verificar se é o mais adequado, criticar a sua execução, no conjunto e nos detalhes;

Analisar cada fenômeno separadamente, fazendo notar os pontos mais em evidência, o máximo e o mínimo, assim como as mudanças mais bruscas;

Investigar se há uma “tendência geral” crescente ou decrescente ou, então, se o fato exposto é estacionário;

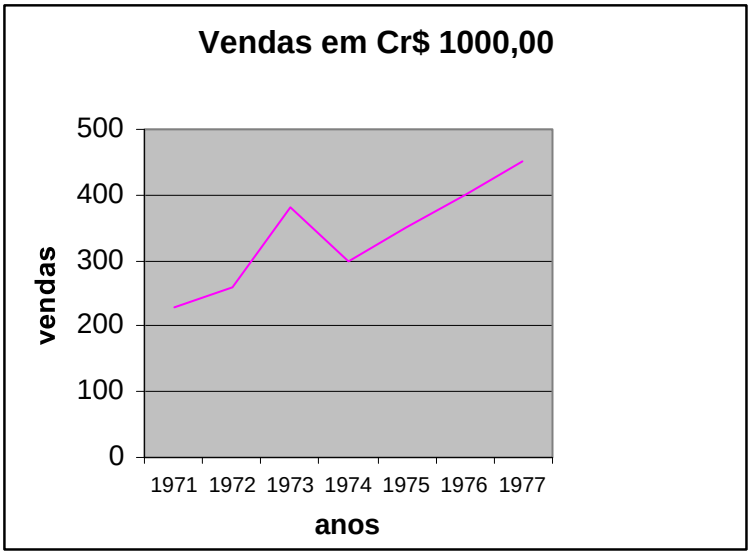
Procurar descobrir a existência de possíveis ciclos periódicos, qual o período aproximado, etc.

Eis os tipos mais comuns de gráficos:

Gráfico em Linhas

Constitui uma aplicação do processo de representação das funções num sistema de coordenadas cartesianas

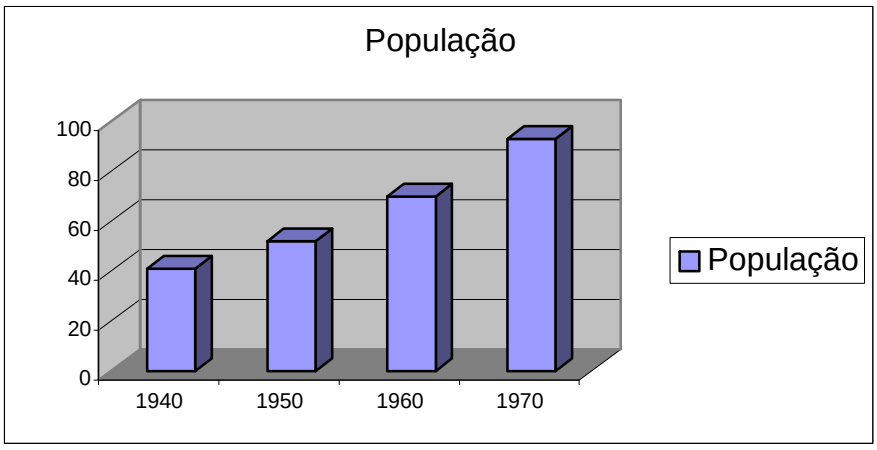
Exemplo: Vendas em Cr$ 1000,00 nos anos de 1971 a 1977 de determinado produto da empresa x.



Fonte: Dados Fictícios.

Gráfico em Colunas

É a representação de uma série por meio de retângulos, dispostos verticalmente. Exemplo: População Brasileira nas décadas de 40 a 70.

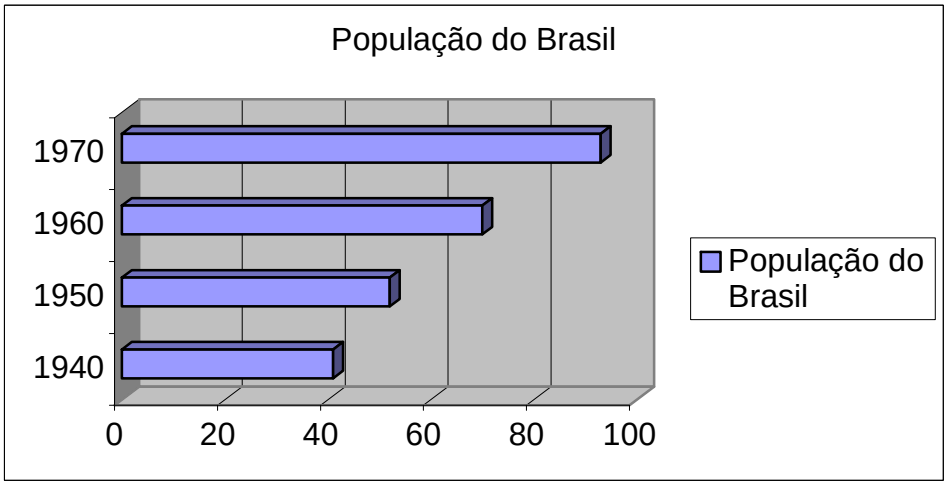


Fonte: Dados Fictícios

Gráfico em Barras

É semelhante ao gráfico em colunas, porém, os retângulos são dispostos horizontalmente.

Exemplo: População Brasileira nas décadas de 40 a 70



Fonte: Dados Fictícios

Gráfico em Setores

É a representação gráfica de uma série estatística em círculo, por meio de setores. É utilizado principalmente quando se pretende comparar cada valor da série com o total.

Exemplo:

Receita (em R$ 1.000.000,00) do Município X de 1975-77

Anos	Receita (em R$ 1.000.000,00)
1975	90
1976	120
1977	
Total	

Fonte: Departamento da Fazenda, Município X.

O total é representado pelo círculo, que fica dividido em tantos setores quantas são as partes. Os setores são tais que suas áreas são respectivamente proporcionais aos dados da série.

Obtemos cada setor por meio de uma regra de três simples e direta, lembrando que o total da série corresponde a 360º.

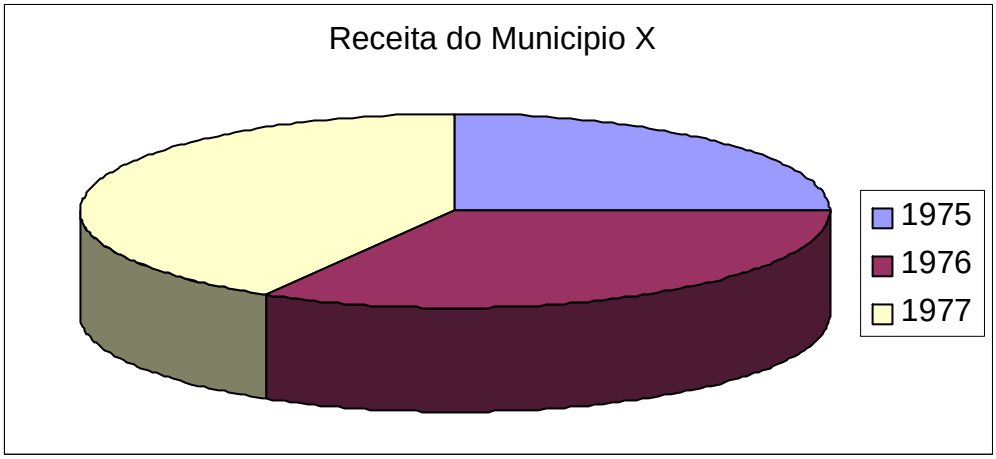
Total _____360º

Parte_____ xº

Para 1975: 360 - 360º
90 - xº
x = 90º

Para 1976: 360 - 360º
120 - xº
x = 120º

Para 1977: 360 - 360º
150 - xº
x = 150º



Fonte**:** Departamento da Fazenda, Município X

Gráfico Polar

É o gráfico ideal para representar séries temporais cíclicas, isto é, séries que apresentam em seu desenvolvimento determinada periodicidade, como, por exemplo, a variação da precipitação pluviométrica ao longo do ano, ou da temperatura ao longo do dia, o consumo de energia elétrica durante o mês ou o ano, etc.

Exemplo:

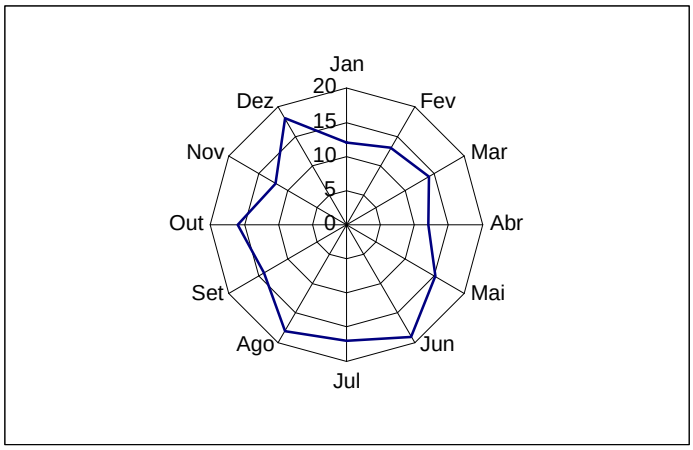
Movimento Mensal de Compras de uma agencia em 1972

Meses	Valores (R$1.000,00)
Janeiro	12
Fevereiro	13
Março	14
Abril	12
Maio	15
Junho	19
Julho	17
Agosto	18
Setembro	14
Outubro	16

Novembro	12
Dezembro	18

Fonte: Departamento financeiro da Agência (dados Fictícios)

Movimento Mensal de Compras de uma agencia em 1972



Fonte: Departamento financeiro da Agência

Amostragem

Veremos quais as técnicas que podemos utilizar para compor uma amostra.

Importância da Amostragem

Na realização de qualquer estudo, quase nunca é possível examinar todos os elementos da população de interesse. Temos usualmente que trabalhar com uma amostra da população. A inferência estatística nos dá elementos para generalizar, de maneira segura, as conclusões obtidas da amostra para a população. Mas, para as inferências serem corretas, é necessário garantir que a amostra seja representativa da população, isto é, a amostra deve possuir as mesmas características básicas da população no que diz respeito ao fenômeno pesquisado.

É errôneo pensar que, caso tivéssemos acesso a todos os elementos da população, seríamos mais precisos. Os erros de coleta e manuseio de um grande

número de dados são maiores do que as imprecisões a que estamos sujeitos quando generalizamos, via inferência, as conclusões de uma amostra bem selecionada.

Em se tratando de amostra, a preocupação central é que ela seja representativa. É preciso que a amostra, ou as amostras que vão ser usadas sejam obtidas por processos adequados.

Assim que decidimos obter informações através de um levantamento amostral, temos imediatamente dois problemas:

- Definir cuidadosamente a população de interesse;
- Selecionar a característica que iremos pesquisar.
- Dados coletados de forma descuidada podem ser tão inúteis que nenhum processamento estatístico consegue salvá-los.

Conceitos Fundamentais

O conceito de **população** é intuitivo; trata-se do conjunto de indivíduos ou objetos que apresentam em comum determinadas características definidas para o estudo.

- **Amostra-** é um subconjunto da população.
- **Amostragem**- são procedimentos para extração de amostras que representem bem a população.
- **Riscos**- é a margem de erro motivado pelo fato de investigarmos parcialmente (amostras) o universo (população).
- **População-alvo**- é a população sobre a qual vamos fazer inferências baseadas na amostra.

Para que possamos fazer inferências válidas sobre a população a partir de uma amostra, é preciso que essa seja representativa. Uma das formas de se conseguir representatividade é fazer com que o processo de escolha da amostra seja, de alguma forma, aleatório. Além disso, a aleatoriedade permite o cálculo de estimativas dos erros envolvidos no processo de inferência.

Quanto à extração dos elementos, as amostras podem ser:

- Com reposição - quando um elemento sorteado puder ser sorteado novamente;
- Sem reposição - quando o elemento sorteado só puder figurar uma única vez na amostra.

Basicamente, existem dois métodos para composição da amostra: probabilístico e não probabilístico (intencional).

- O método de amostragem probabilística exige que cada elemento da população possua determinada probabilidade de ser selecionado. Normalmente, possuem a mesma probabilidade. Assim, se N for o tamanho da população, a probabilidade de cada elemento será 1/N. Somente com base em amostragens probabilísticas pode-se realizar inferências sobre a população, a partir dos parâmetros estudados na amostra. São elas:
- Amostragem Aleatória Simples;
- Amostragem Aleatória Estratificada;
- Amostragem Sistemática;
- Amostragem por Conglomerado.

Por serem as principais técnicas estudas, serão mais detalhadamente exploradas no item 2.3.

- Os métodos não probabilísticos são amostragens em que há uma escolha deliberada dos elementos que compõem a amostra. Não se pode generalizar os resultados das pesquisas para a população, uma vez que as amostras não probabilísticas não garantem a representatividade da população. São elas:
- Amostragem Acidental;
- Amostragem Intencional;
- Amostragem por Quotas.

Amostragem Acidental **-** É formada por elementos que vão aparecendo, que são possíveis de se obter até completar o número de elementos da amostra.

Ex: Pesquisa de opinião, em que os entrevistados são acidentalmente escolhidos.

Amostragem Intencional - É formada por elementos escolhidos por determinado critério, ou seja, escolhe-se intencionalmente um grupo de elementos que irão compor a amostra.

Amostragem por Cotas - Classificação da população em termos de propriedades que se sabe serem relevantes para a característica a ser estudada. Determinação da proporção da população para cada característica com base na constituição conhecida, ou estimada, da população. Fixação de quotas para cada observador, ou entrevistador, a quem tocará a responsabilidade de selecionar interlocutores ou entrevistados, de modo que a amostra total observada, ou entrevistada, contenha a proporção de cada classe.

Amostragem Aleatória Simples

A amostragem aleatória simples é um processo para selecionar amostras de tamanho “n” dentre as “N” unidades em que foi dividida a população. Sendo a amostragem realizada sem reposição, que é o caso mais comum, existem (N,n) possíveis amostras, todas igualmente prováveis. As amostras aleatórias podem ser escolhidas por diversos métodos, inclusive por tabelas de números aleatórios (TNA) e de computadores para gerar números aleatórios Na prática, a amostra aleatória simples é escolhida unidade por unidade. As unidades da população são numeradas de 1 a N. Em seguida, escolhe-se, na tabela de números aleatórios (TNA), (ou por computador) n números compreendidos entre 1 e N. Esse processo é equivalente a um sorteio no qual se colocam todos os números misturados dentro de uma urna.

As unidades correspondentes aos números escolhidos formarão a amostra.

Amostragem Aleatória Estratificada

Uma amostra estratificada é obtida separando-se as unidades da população em grupos não superpostos chamados estratos, e selecionando-se independentemente uma amostra aleatória simples de cada estrato. Existem dois tipos de amostragem estratificada:

- De igual tamanho;
- Proporcional.

No primeiro tipo, sorteia-se igual número de elementos em cada estrato. Esse processo é utilizado quando o número de elementos por estrato for aproximadamente o mesmo.

No outro caso, utiliza-se a amostragem estratificada proporcional, cujo processo de calcular o número de amostras por estrato é:

N → Nº de unidades da população

n → Nº de unidades das amostras

N_a →Nº de unidades do estrato A

n_a → Nº de amostras de A

$$\frac{N_a}{N} = \frac{n_a}{n} \rightarrow n_a = \frac{n}{N}.N_a$$

Exemplo:

Supondo, no exemplo anterior, que, dos noventa alunos, 54 sejam meninos e 36 sejam meninas, vamos obter uma amostra proporcional estratificada de 10%.

Resolução:

- São, portanto, dois estratos (sexo masculino e feminino) e queremos uma amostra de 10% da população;
- Calcula-se o número de amostras de cada estrato.

Sexo	População	10%	Número de amostras
M	54	5,4	5
F	36	3,6	4
Total	90	9,0	9

- Numeramos os alunos de 01 a 90, sendo que de 01 a 54 correspondem meninos e de 55 a 90, meninas. O próximo passo é o mesmo do exemplo anterior.

Amostragem por Conglomerado

Uma amostra por conglomerado é uma amostra aleatória simples na qual cada unidade de amostragem é um grupo, ou um conglomerado de elementos.

O primeiro passo na amostragem por conglomerado é especificar conglomerados apropriados. Os elementos em um conglomerado tendem a ter características similares, portanto, o fato de novas medidas serem tomadas num conglomerado não implica necessariamente aumento de informação sobre o parâmetro populacional. Como regra geral, o número de elementos num conglomerado deverá ser pequeno em relação ao tamanho da população e o número de conglomerados deverá ser razoavelmente grande.

Na amostragem por conglomerado a população é dividida em grupos. E selecionam-se amostras aleatórias simples de grupos e, então, todos os itens dos grupos (conglomerados) selecionados farão parte da amostra.

Exemplo:

Em um levantamento da população de uma cidade, podemos dispor do mapa indicando cada quarteirão e não dispor de uma relação atualizada dos seus moradores. Pode-se, então, colher uma amostra dos quarteirões e fazer a contagem completa de todos os que residem naqueles quarteirões sorteados.

Amostragem Sistemática

Quando os elementos da população já se encontram ordenados, não há necessidade de se construir o sistema de referência. Nesses casos, a seleção dos elementos que constituirão a amostra pode ser por um sistema imposto pelo pesquisador.

Em geral, para se obter uma amostra sistemática de n elementos de uma população de tamanho N, K deve ser menor ou igual a N/n. Não é possível determinar K, precisamente, quando o tamanho da população é desconhecido, mas pode-se supor um valor de k de tal modo que seja possível obter uma amostra de tamanho n. Em vez da amostragem aleatória simples, pode-se empregar a amostragem sistemática pelas seguintes razões:

- a amostragem sistemática é mais fácil de se executar e, por isso, está menos sujeita a erros do entrevistador do que aqueles que acontecem na aleatória simples; a amostragem sistemática frequentemente proporciona mais informações por custo unitário do que a aleatória simples.

  Diretrizes para calcular as amostras:

  1º - Estabelecer o intervalo de amostragem K:

  $$K = \frac{N}{n}$$

  OBS: Para valores de K=N/n , arredondar para o valor inteiro menor.

  2º - Iniciar aleatoriamente a composição da amostra.

  b → inicio (nº de ordem inicial sorteado na TNA).

  OBS: **0 < b ≤ K**

  3º - Composição da Amostra:

  1º item →b

  2º item →b + K

  3º item →b + 2k

  Exemplo:

  1 – Suponhamos uma rua contendo quinhentos prédios, dos quais desejamos obter uma amostra formada de vinte prédios. (TNA, 3ªLinha e 5ª Coluna)

  Solução:

a) Calcular K (intervalo de amostragem)
b) K=500/20, K=25
c) b= 12 (valor encontrado na TNA)
d) Composição da amostra

   1º item → 12

   2º item → 12 + 25 = 37

   3º item → 12 + 2*25 = 62

   20º item → 12 +19*25 = 487

Tabela de Números Aleatórios

COLUNA	1	5	9	13	17	21	25	29	33	37	41	45	49
LINHA													
01	9486	9821	6074	1432	0995	0157	0071	9871	6678	0140	9522	0995	1735
02	3155	9878	3359	8244	8952	0084	1558	4775	1699	1652	2555	4765	2709
03	6136	2824	6030	4256	3870	5725	2204	5318	8337	3867	6184	2018	3522
04	7249	9182	8669	7423	1768	8147	7285	8390	9134	9863	9486	9821	6074
05	0071	9871	6678	0140	9522	0995	1735	1248	9807	1910	3155	9878	3359
06	1558	4775	1699	1652	2555	4765	2709	0561	4397	1135	6136	2824	6030
07	2204	5318	8337	3867	6184	2018	3522	0941	5569	5800	7249	9182	8669
08	7285	8390	9134	9863	9486	9821	6074	1432	0995	0157	0071	9871	6678
09	1735	1248	9807	1910	3155	9878	3359	8244	8952	0084	1558	4775	1699
10	2709	0561	4397	1135	6136	2824	6030	4256	3870	5725	2204	5318	8337
11	3522	0941	5569	5800	7249	9182	8669	7423	1768	8147	7285	8390	9134
12	6074	1432	0995	0157	0071	9871	6678	0140	9522	0995	1735	1248	9807
13	3359	8244	8952	0084	1558	4775	1699	1652	2555	4765	2709	0561	4397
14	6030	4256	3870	5725	2204	2318	8337	3867	6184	2018	3522	0941	5569
15	8669	7423	1768	8147	7285	8390	9134	9863	9486	9821	6074	1432	0995
16	6678	0140	9522	0995	1735	1248	9807	1910	3155	9878	3359	8244	8952
17	1699	1652	2555	4765	2709	0561	4397	1135	6136	2824	6030	4256	3870
18	8337	3867	6184	2018	3522	0941	5569	5800	7249	9182	8669	7423	1768
19	9134	9863	9486	9821	6074	1432	0995	0157	0071	9871	6678	0140	9522
20	9807	1910	3155	9878	3359	8244	8952	0084	1558	4775	1699	1652	2555
21	4397	1135	6136	2824	6030	4256	3870	5725	2204	5318	8337	3867	6184
22	5569	5800	7249	9182	8669	7423	1768	8147	7285	8390	9134	9863	8486
23	0995	0157	0071	9871	6678	0140	9522	0995	1735	1248	9807	1910	3155
24	8952	0084	1558	4775	1699	1652	2555	4765	2709	0568	4397	1135	6136
25	3870	5725	2204	5318	8337	3867	6184	2018	3522	0941	5569	5800	7249
26	7425	3566	6151	4731	6489	2491	2765	8525	7849	1488	8833	2597	1333
27	8961	8175	0879	6945	8029	9119	5990	1063	9444	8320	1740	6131	9907
28	3298	6173	1741	3874	9321	3748	7507	0170	0568	9112	1275	0924	3054
29	2276	4898	2394	1098	4063	5393	0226	8144	4778	7471	1764	4939	8063
30	9557	8114	1576	9767	1486	7161	5606	6295	3503	5050	9549	2500	9666
31	8650	1920	2533	7755	5324	3731	3414	2153	3815	0626	5718	8679	6801
32	2885	8101	1467	0080	7962	5999	9562	5819	1562	6793	2065	0239	8253
33	1841	8626	0344	4344	7446	0867	6157	8935	4413	2363	7187	8980	2488
34	4638	8030	0018	7760	9819	4276	0650	3516	5159	9236	3257	1694	7157
35	1320	7033	1218	5605	4206	2878	0230	1740	4553	8729	5827	7176	8703
36	1488	5803	6790	9368	0465	4819	0065	7633	3950	2109	7027	5824	5057
37	4353	4347	8565	2231	8789	4231	2585	0157	2037	7835	1320	8999	9181
38	7816	5817	9764	8789	7387	2172	0896	1038	6047	9539	3510	1343	8098
39	8600	9738	5415	8426	7152	8705	5829	0164	8330	9152	6045	8129	2293
40	1057	1550	8773	3003	4302	4034	2478	1078	0429	7189	0778	3260	5969

Distribuição de Frequência

Vamos considerar, o estudo detalhado da distribuição de frequência, que é a forma pela qual podemos descrever os dados estatísticos resultantes de variáveis quantitativas. São objetivos:

- Compor uma distribuição de frequência com ou sem intervalos de classe;
- Determinar o quadro de frequências, eles são úteis para condensar grandes conjuntos de dados, facilitando o sua utilização;
- Representar uma distribuição de frequência através de histograma, polígono e ogiva.

Conceitos

Ao analisarmos um conjunto de dados, devemos determinar se temos uma **amostra** ou uma população. Essa determinação afetará não somente os métodos utilizados, mas também as conclusões, pois se estamos trabalhando com uma amostra os resultados encontrados são estimativas da população.

Nem sempre é possível compreender o significado contido numa amostragem por simples inspeção visual dos dados numéricos coletados. Entretanto, entendemos que o sucesso de uma decisão dependerá da nossa habilidade em compreender as informações contidas nesses dados. O objetivo deste estudo é mostrar a organização, apresentação e análise gráfica de uma série de dados, matéria prima das **distribuições de frequências** e dos histogramas. Frequência de uma observação é o número de repetições dessa observação, ou seja, quantas vezes determinado fenômeno acontece.

Os dados podem ser classificados como:

- **Dados brutos** – são os dados originais, que ainda não se encontram prontos para análise, por não estarem numericamente organizados. (Também são conhecidos como Tabela Primitiva).

Exemplo: Número mensal de aparelhos defeituosos na Empresa X.

	J	F	M	A	M	J	J	A	S	O	N	D
1995	6	2	5	1	0	3	2	1	3	5	5	3
1996	5	4	2	1	3	4	1	4	5	4	0	1
1997	3	1	2	4	3	1	4	1	0	3	0	2
1998	2	2	0	3	1	4	2	0	1	1	5	2

- **Rol –** são os dados brutos, organizados em ordem crescente ou decrescente.

  Exemplo: Considerando o exemplo anterior temos:

0	0	0	0	0	0	1	1	1	1	1	1
1	1	1	1	1	2	2	2	2	2	2	2
2	2	3	3	3	3	3	3	3	3	4	4
4	4	4	4	4	5	5	5	5	5	5	6

- **Dados discretos** – a variável é discreta quando assume valores em pontos da reta real.

  Exemplo: número de erros em um livro: 0,1,2,3,4,.... número de filhos de vários casais: 1,2,3,4,.....

  quantidade de acidentes em determinada rodovia: 4,10,12,15,....

- **Dados contínuos** – a variável pode assumir, teoricamente, qualquer valor em certo intervalo da reta real.

  Exemplo: peso de alunos: 55,5 kg; 61,0kg; 63,4 kg; 68,1 kg.....

  distância entre cidades: 35,5 km; 48,6 km; 100,10 km; ....

- **Dados Tabelados não agrupados em classes** – os valores da variável aparecem individualmente.

  Exemplo, considerando os dados da tabela anterior:

Nº de aparelhos com defeitos	Nº de meses
0	06
1	11
2	09
3	08
4	08
5	05
6	01
Total	48

- **Dados Tabelados agrupados em classes** - os valores da variável não aparecem individualmente, mas agrupados em classes.

Notas	Nº de alunos
0 |--- 20	020
20 |--- 40	065
40 |--- 60	230
60 |--- 80	160
80 |--- 100	025
Total	580

Elementos de uma distribuição de frequência: amplitude total, limites de classe, amplitude do intervalo de classe, ponto médio da classe, frequência absoluta, relativa e acumulada

Amplitude total (A) - é a diferença entre o maior e o menor número do rol.

Exemplo: Estatura de 40 alunos do Colégio A em cm. (Dados ordenados em ordem crescente, por colunas)

150	154	155	157	160	161	162	164	166	169
151	155	156	158	160	161	162	164	167	170
152	155	156	158	160	161	163	164	168	172
153	155	156	160	160	161	163	165	168	173

A = 173 – 150 = 23

Número de classes (K) e Classe (i) – não existe regra fixa para se determinar o número de classes. Podemos utilizar:

- A Regra de Sturges, que nos dá o número de classes em função do número de valores da variável:

  K=1+3,3.log n, onde n é o número de itens que compõe a amostra;

- Ou

$$K = 5 \text{ para } n \leq 25 \text{ e } k \cong \sqrt{n}, \text{ para } n > 25.$$

Exemplo: considerando o exemplo anterior n=40

- Pela formula de Sturges: K= 1+3,3log40 = 6,28 → K=6
- Adotando K = n , temos $\sqrt{k} = 40 = 6,3 \rightarrow K=6$

Amplitude de um intervalo de classe (h) – ou simplesmente intervalo de classe é a medida do intervalo que define a classe.

h = A / K

Exemplo, considerando o exemplo anterior:

H = 23/ 6 = 3,83 → h = 4

Limites de Classe – denominamos limites de classe os extremos de cada classe. Assim temos:

- limite inferior (linf) e
- limite superior (Lsup)

Observação: Vamos trabalhar com intervalos fechados à esquerda e abertos à direita; isso significa que valores iguais ou superiores ao limite inferior são considerados nessa classe e valores iguais e/ou superiores ao limite superior são considerados na classe abaixo.

Exemplo: Do exemplo anterior, temos:

i	*Classes*	*n*				
1	150 ⊢ 154	4				
2	154 ⊢ 158	9				
3	158 ⊢ 162	11				
4	162 ⊢ 166	8				
5	166 ⊢ 170	5				
6	170 ⊢ 174	3				
	Σ	40				

Na segunda classe, temos:

- L_2=158
- l_2 = 154

Ponto Médio da Classe - (x_i) – é, como o próprio nome indica, o ponto que divide o intervalo de classe em duas partes iguais. Para obtermos o ponto médio de uma classe, calculamos:

$$x_i = \frac{l_{inf} + L_{sup}}{2}$$

Exemplo: considerando a segunda classe do exemplo anterior, temos:

$x_2 = \frac{154+158}{2} = 156 \rightarrow x_2 = 156$

Frequências

- **Frequências simples ou absoluta da classe i (n_i)** - são os valores que realmente representam o número de dados de cada classe. A soma das frequências simples é igual ao número total dos dados.

  Exemplo: considerando a segunda classe do exemplo anterior, temos: $n_2 = 9$.

- **Frequências relativas (f_i)** – são os valores das razões entre as frequências simples e o número total de dados.

$$f_i = \frac{n_i}{n}$$

  Exemplo: considerando a segunda classe do exemplo anterior, temos:

  $f_2 = 9/40 = 0{,}225$.

  Obs: as frequências relativas permitem a análise ou facilitam as comparações.

- **Frequência acumulada (Ni)** – é o total das frequências de todos os valores inferiores ao limite superior do intervalo de uma dada classe:

  Exemplo: Considerando a frequência acumulada da quarta classe (N_4), temos:

  $N_4 = n_1+n_2+n_3+n_4 = 4+9+11+8 = 32$

- **Frequência acumulada relativa (F_i)** – é a frequência acumulada da classe, dividida pela frequência total da distribuição.

$$F_i = \frac{N_i}{n}$$

  Exemplo: para o exemplo anterior, $F_4 = 0{,}8$.

NOTA – Usualmente, denominamos**:**

- **Frequência relativa acumulada crescente** da classe i – F_i.

Obs: F_i pode ser entendido como sendo a percentagem de observações abaixo do limite superior da classe i.

- **Frequência relativa acumulada decrescente** da classe i – F'_i.

Obs: F'_i e a porcentagem de observações acima do limite inferior da classe i.

Regras Gerais para a elaboração de uma distribuição de frequência

Os principais estágios na construção de uma distribuição de frequência para dados amostrais são:

1. Encontrar a amplitude total do conjunto de valores observados;
2. Escolher o número de classes;

$$K = 1 + 3{,}3 \log n \quad \text{ou} \quad k = \sqrt{n}$$

3. Determinar a amplitude do intervalo de classe;

$$h = \frac{A}{k}$$

4. Determinar os limites de classe;
5. Construir a tabela de frequências.

Exemplo:

Calcule as frequências e o ponto médio dos dados abaixo:

Alturas de 50 estudantes do sexo masculino da Univesidade XYZ

33	35	35	39	41	41	42	45	47	48
50	52	53	54	55	55	57	59	60	60
61	64	65	65	65	66	66	66	67	68
69	71	73	73	74	74	76	77	77	78
80	81	84	85	85	88	89	91	94	97

Solução: Amplitude : A = 97-33 = 64

Número de Classes : $K = \sqrt{50} \cong 7$

Intervalo de classe : h =64/7 »10

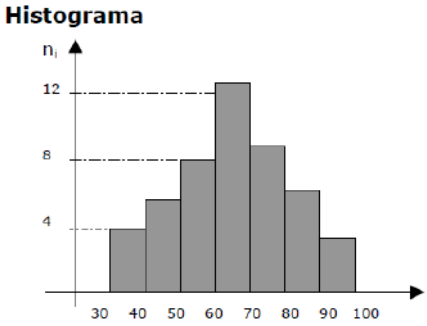
i	*Classes*	*n*	N_i	f_i	F_i	x_j
1	30 ⊢ 40	4	4	0,08	0,08	35
2	40 ⊢ 50	6	10	0,12	0,20	45
3	50 ⊢ 60	8	18	0,16	0,36	55
4	60 ⊢ 70	13	31	0,26	0,62	65
5	70 ⊢ 80	9	40	0,18	0,80	75
6	80 ⊢ 90	7	47	0,14	0,94	85
7	90 ⊢100	3	50	0,06	1,0	95
	Σ	50		1		

Gráficos representativos de uma distribuição de frequência: histograma, polígono de frequência e ogiva

Uma distribuição de frequência pode ser representativa graficamente pelo histograma, pelo polígono de frequência e pelo polígono de frequência acumulada (Ogiva de Galton).

Histograma – é formado por um conjunto de retângulos justapostos, cujas bases se localizam sobre o eixo horizontal, de tal modo que seus pontos médios coincidem com os pontos médios dos intervalos de classe.

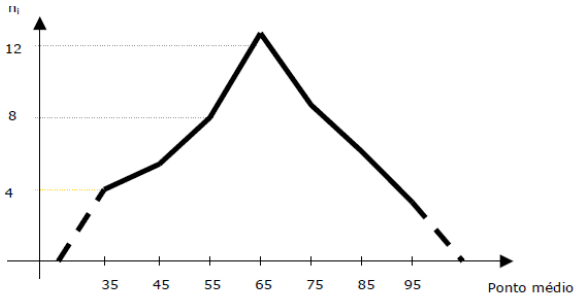
- As larguras dos retângulos são iguais às amplitudes dos intervalos de classe.
- As alturas dos retângulos devem ser proporcionais às frequências das classes, sendo igual a amplitude dos intervalos.



Polígono de frequência – é um gráfico em linha, sendo as frequências marcadas sobre perpendiculares ao eixo horizontal, levantada pelos pontos médios dos intervalos de classe.

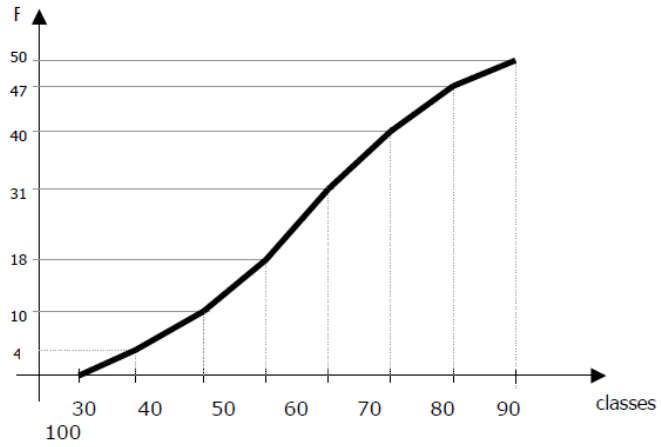
Para realmente obtermos um polígono (linha fechada), devemos completar a figura, ligando os extremos da linha obtida aos pontos médios da classe anterior à primeira e da posterior à última, da distribuição.

Polígono de Frequência



Polígono de frequência acumulada – é traçado marcando-se as frequências acumuladas sobre perpendiculares ao eixo horizontal, levantadas nos pontos correspondentes aos limites superiores dos intervalos de classe.

Polígono de Frequência Acumulada



Observação: Uma distribuição de frequência sem intervalos de classe é representada graficamente por um diagrama onde cada valor da variável é representado por um segmento de reta vertical e de comprimento proporcional à respectiva frequência.

Medidas de Posição

Veremos as tendências características de cada distribuição, destacando as medidas de posição central, que recebem tal denominação pelo fato de os dados observados tenderem, em geral, a se agrupar em torno dos valores centrais.

Nas seções anteriores, vimos a sintetização dos dados sob a forma de tabelas, gráficos e distribuições de frequências. Agora, vamos destacar o cálculo das medidas que possibilitam localizar a maior concentração de valores de uma dada distribuição, isto é, se ela se localiza no início, no meio ou no final, ou, ainda, se há uma distribuição

por igual. Tais medidas possibilitam comparações de séries de dados entre si pelo confronto desses números.

No entanto, para ressaltar as tendências características de cada distribuição, isoladamente, ou em confronto com outras, necessitamos introduzir os elementos típicos da distribuição, que são:

- Medidas de posição;
- Medidas de variabilidade ou dispersão; • Medidas de assimetria;
- Medidas de curtose.

As medidas de posição mais importantes são as medidas de tendência central, que destacamos:

- A média aritmética;
- A mediana;
- A moda.

As outras medidas de posição são as separatrizes, que englobam:

- A mediana;
- Os quartis;
- Os decis;
- Os percentis.

Primeiramente, vamos estudar as principais medidas de tendência central, depois veremos as separatrizes e, as medidas de Dispersão, Assimetria e Curtose.

Média aritmética simples e ponderada e suas propriedades

É o quociente da divisão da soma dos valores da variável pelo número deles. A média (aritmética) é, de modo geral, a mais importante de todas as medidas descritivas.

Dados não tabelados

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

xi : valor observado

n : número total de observações

Exemplo: Suponha que o tempo de vida útil de 10 aparelhos de telefone são:
10 29 26 28 15 23 17 25 0 20.

Qual a média de vida útil destes aparelhos?

Solução:

$$\sum X = 10 + 29 + 26 + 28 + 15 + 23 + 17 + 25 + 0 + 20 = 193$$

$$\bar{x} = \frac{\sum x}{n} = \frac{193}{10} = 19{,}3 \ ,$$

Portanto média de vida útil dos aparelhos são 19,3 anos.

Dados Tabelados

- Sem intervalo de Classe

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \times n_i}{n}$$

X_i : valor observado

n_i: n° de observações por classe

n: nº de observações totais

- Com intervalo de Classe

$$\overline{X} = \frac{\sum_{i=1}^{n} \overline{x_i} \times n_i}{n}$$

$\overline{xi}$: ponto médio da classe

n_i : n° de observações

Às vezes, a média pode ser um número diferente de todos os da série de dados que ela representa.

Moda (M_o): Dados agrupados e não agrupados em classes

É o valor que ocorre com maior frequência em um conjunto de dados, e que é denominado valor modal. Baseado nesse contexto, um conjunto de dados pode apresentar mais de uma moda. Nesse caso, dizemos ser multimodais; caso contrário, quando não existe um valor predominante, dizemos que é amodal.

Dados não tabelados: o valor modal é o predominante na distribuição.

Exemplo: Nos valores abaixo, qual o valor modal?

3	4	4	5	6	7	8	9	9	9	10	11	12	13

Solução: M_o= 9

Dados tabelados

- Sem intervalo de Classe: O valor modal é o valor que possuir maior frequência.

Exemplo:

Classes	0	1	2	3	4	5	6	Σ

Solução: o valor predominante é o número 1, que ocorreu 11 vezes.

Temos, portanto, M_o=1.

- Com intervalo de classe: tratando-se de dados agrupados em classe, a moda não é percebida tão facilmente como nos casos anteriores. Para calcular o valor modal nesses casos, podemos utilizar os seguintes processos:

1º Processo: Fórmula de Czuber

$$Mo = l_{\inf} + h_{Mo} \frac{n_{mo} - n_{ant}}{(n_{mo} - n_{ant}) + (n_{mo} - n_{post})}$$

Onde constatamos:

Classe Modal: Classe de maior frequência

n_{mo}: frequência simples da classe modal

n_{ant} : frequência simples anterior à classe modal

n_{post} : frequência simples posterior à classe modal

l_{inf}: limite inferior da classe modal

h_{Mo}: intervalo de classe modal

2º Processo: Fórmula de Pearson

$$M_o = 3Md - 2\overline{X}$$

onde constatamos:

M_d = Mediana

$\overline{X}$ = Média

Exemplo - Determinar a moda para a distribuição:

Classes	0|--- 1	1|--- 2	2|--- 3	3|--- 4	4|--- 5	Σ
n_i	3	10	17	8	5	43

Solução: Utilizando a fórmula de Czuber

- a classe modal é a classe 2|----3
- $l_{inf} = 2$
- $h_{Mo} = 1$
- $n_{Mo} = 17$
- $n_{ant} = 10$
- $n_{post} = 8$

Substituindo esses valores na fórmula, encontramos: $M_o = 2,44$

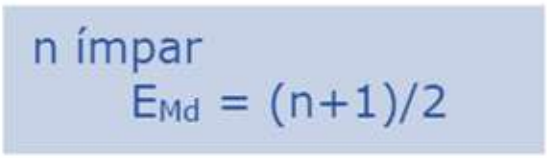
Mediana (M_d): Dados agrupados e não agrupados em classes

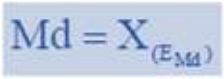
A mediana é uma medida de posição. É, também, uma separatriz, pois divide o conjunto em duas partes iguais, com o mesmo número de elementos. O valor da mediana encontra-se no centro da série estatística organizada, de tal forma que o número de elementos situados antes desse valor (mediana) é igual ao número de elementos que se encontram após esse mesmo valor (mediana).

Dados não tabelados

- Para uma série com número ímpar de itens: a mediana corresponde ao valor central.

EMd - elemento mediano: indica a posição da mediana.





A mediana será o termo de ordem (n+1)/2.

- Para uma série com número par de itens: não há termo central único, mas, sim, dois termos centrais. A mediana será dada por:

n par

$$E_{Md} = n/2$$

$$Md = \frac{X_{(E_{Md})} + X_{(E_{Md}+1)}}{2}$$

A mediana será a média aritmética entre os termos centrais.

Dados tabelados:

Neste caso, o problema consiste em determinar o ponto do intervalo em que está compreendida a mediana.

- Sem intervalo de classe: devemos, primeiro, obter a localização da mediana na série através do termo de ordem:

$$E_{MD} = \frac{n}{2}$$

Uma vez localizada a posição da mediana, devemos verificar o valor numérico da variável correspondente a essa posição.

- Com intervalo de classe: localizada a classe mediana, calculamos o valor da mediana pela fórmula:

$$Md = l_{inf} + h_{Md} \frac{\sum \frac{n}{2} - N_{ant}}{n_{Md}}$$

onde temos:

l_{inf}: limite inferior da classe mediana.

h_{Mo}: intervalo de classe mediana.

n_{Md} : frequência simples absoluta na classe mediana.

N_{ant} : frequência acumulada absoluta anterior à classe mediana.

Classe Mediana: classe onde está o elemento mediano.

Média Geométrica (M_G): Dados agrupados e não agrupados em classes

Dados não tabelados

A média geométrica de um conjunto de N números x_1, x_2,x_3,.....x_n é a raiz de ordem N do produto desses números:

$$M_G = \sqrt[N]{x_1 .. x_2 .\ x_3 ..... x_n}$$

Dados agrupados

- Sem intervalo de classe

$$M_G = \sqrt[N]{x_1^{n_1} x_2^{n_2} x_3^{n_3} ....x_n^{n_n}}$$

X_i : valor observado
n_i : n° de observações da classe

- Com intervalo de classe

$$M_G = \sqrt[N]{\bar{x}_1^{n_1} \bar{x}_2^{n_2} \bar{x}_3^{n_3} ....\bar{x}_n^{n_n}}$$

$\bar{x}_i$: ponto médio
n_i : n° de observações

Média Harmônica (M_h): Dados agrupados e não agrupados em classes

Sejam x_1, x_2, x_3,.....x_n, valores de x, associados às frequências absolutas n_1, n_2, n_3,.....n_n, respectivamente.

A média harmônica de X é definida por:

$$Mh = \frac{n}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \frac{n_3}{x_3} + ... + \frac{n_n}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{n_i}{x_i}}$$

- Para dados não agrupados n = 1.
- Para dados agrupados sem intervalo de classe x_i é o valor da variável.
- Para dados agrupados com intervalo de classe x_i é o ponto médio da classe.

Separatrizes: Quartis, Decis e Percentis

São valores que ocupam determinados lugares de uma distribuição de frequência.

Podemos classificá-las em:

Quartis: dividem a distribuição em 4 partes iguais

Qi = quartil i=1,2,3

- Q1 = 1º quartil, valor situado de tal modo na série que uma quarta parte (25%) dos dados é menor que ele e as três quartas partes restantes (75%) são maiores.
- Q2 = 2º quartil, evidentemente, coincide com a Mediana (Q2 = Md).
- Q3 = 3º quartil, valor situado de tal modo que as três quartas partes (75 %) dos termos são menores que ele e uma quarta parte 25 % é maior.

$$Q_1 = l_{Q_1} + \frac{\left(\frac{n}{4} - N_{ant}\right)}{n_{Q_1}}.h \qquad Q_3 = l_{Q_3} + \frac{\left(\frac{3n}{4} - N_{ant}\right)}{n_{Q_3}}.h$$

Onde temos:

l_{inf} : limite inferior da classe do quartil considerado

h_Q: intervalo de classe do quartil considerado

n_Q : frequência simples absoluta do quartil considerado

N_{ant} : frequência acumulada anterior à classe do quartil considerado

Decis: dividem a distribuição em 10 partes iguais

D_i = decil i=1,2,3, …, 9

1º Passo: Calcula-se em que K = 1,2,3,.....,9;

2º Passo: Identifica-se a classe D_K , pela N_{ac} ;

3º Passo: Aplica-se a fórmula:

$$D_K = l_{D_K} + \frac{\left(\frac{KN}{10} - N_{(ant)}\right)}{n_{D_K}} h$$

l_{DK}: limite inferior da Classe Dk

N : tamanho da amostra

h : amplitude da classe

n_{DK}: frequência da classe

$N_{(ant)}$: frequência acumulada da classe

Percentis: dividem a distribuição em 100 partes iguais.

P_i = centil i=1,2,3, …, 99

1º Passo: Calcula-se em que K = 1,2,3,4,.....98,99

2º Passo: Pela N_{ac} identifica-se a classe Pi

3º Passo: Aplica-se a fórmula

$$P_K = l_{P_K} + \frac{\left(\frac{KN}{100} - N_{(ant)}\right)}{n_{P_K}} h$$

l_{PK}: limite inferior da classe em que, K = 1,2,3,...,98,99.

N : tamanho da amostra

$N_{(ant)}$: frequência acumulada anterior à classe

h : amplitude da classe

n_{PK}: frequência da classe

Exemplo:

1- Num acampamento infantil, foram obtidas as seguintes estaturas:

2-

Estaturas	120|--- 128	128|---136	136|--- 144	144|--- 152	152|--- 160
frequencia	6	12	16	13	7

Calcule:

a) O 1º Quartil (Q_1);

b) o 4º Decil (D_4); Solução:

Primeiro vamos estruturar a tabela de distribuição de Frequências, como estamos trabalhando com intervalos de classe, temos que calcular os pontos médios de cada classe. Depois iremos utilizar as fórmulas para cada item que queremos calcular.

i	Estaturas (cm)	n	N	X_i (Ponto médio)	
1	120|--- 128	6	6	124	
2	128 |--- 136	12	18	132	
3	136 |--- 144	16	34	138	
4	144 |--- 152	13	47	148	
5	152 |--- 160	7	54	156	
	Total	54			

a) Calculo de Q_1,

Para calcular Q_1, temos que primeiro identificar a classe que está o valor, para isto consideramos:

$N = 54 = 13,5$, que vamos neste momento arredondar para 14, pela frequência
4 4

acumulada procuramos a classe que encontra o 14º elemento, que é a 2ª classe com limites de 128 |--- 136.

Agora usamos a formula para calcular Q_1

$$Q_1 = l_{Q_1} + \frac{\left(\frac{n}{4} - N_{ant}\right)}{n_{Q_1}}.h$$

Onde:

l_{inf} : limite inferior da classe do quartil considerado = 128

h_Q: intervalo de classe do quartil considerado = 8

n_Q : frequência simples absoluta do quartil considerado = 12

N_{ant} : frequência acumulada anterior à classe do quartil considerado =6

Substituindo estes valores na expressão acima temos

$Q_1 = 128 + \frac{(13{,}5 - 6)}{12} * 8 = 133$

b) Calculo de D_4

Primeiro identificamos a classe que esta o valor;

KN = 4x54 = 21,6 = 27, através da frequência acumulada identificamos a clas-
10 10

se que encontra o 27º , que é a 3ª classe com limites de 136 |--- 144

Agora usamos a fórmula

$$D_K = l_{D_K} + \frac{\left(\frac{KN}{10} - N_{(ant)}\right)}{n_{D_K}}h$$

onde:

lDK: limite inferior da Classe Dk = 136

N : tamanho da amostra = 54

h : amplitude da classe = 8

n_{DK}: frequência da classe = 16

$N_{(ant)}$: frequência acumulada da classe anterior = 18

Substituindo estes valores na expressão acima temos:

$$D_4 = 136 + \frac{(21{,}6 - 18)}{16} * 8 = 137{,}8$$

Medidas de Dispersão

A interpretação de dados estatísticos exige que se realize um número maior de estudos, além das medidas de posição. veremos que as medidas de dispersão servem para verificar a representatividade das medidas de posição, que é o nosso principal objetivo.

Dispersão

São medidas estatísticas utilizadas para avaliar o grau de variabilidade, ou dispersão, dos valores em torno da média. Servem para medir a representatividade da média.

Absoluta

- Amplitude (A)
- Variância (S2)
- Desvio padrão (S)

Relativa

• Coeficiente de Variação (CV)

Amplitude Total (A)

É a diferença entre o maior e o menor dos valores da série. A utilização da amplitude total como medida de dispersão é muito limitada, pois sendo uma medida que depende apenas dos valores externos, é instável, não sendo afetada pela dispersão dos valores internos.

Variância (S^2)

A variância leva em consideração os valores extremos e os valores intermediários, isto é, expressa melhor os resultados obtidos. A variância relaciona os desvios em torno da média, ou, mais especificamente, é a média aritmética dos quadrados dos desvios.

Dados Brutos Dados Agrupados

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad S^2 = \frac{\sum x_i^2 n_i}{n-1} - \left(\frac{\sum x_i n_i}{n-1}\right)^2$$

Onde temos que:

- Para dados agrupados sem intervalo de classe, x_i é o valor da variável.
- Para dados agrupados com intervalo de classe, x_i é o ponto médio da classe.

Para dados amostrais, o denominador é n-1**;** para dados populacionais, usamos no denominador o valor de n**.**

Sendo a variância calculada a partir dos quadrados dos desvios, ela é um número em unidade quadrada em relação à variável em questão, o que, sob o ponto de vista prático, é um inconveniente; por isso, tem pouca utilidade na estatística descritiva, mas é extremamente importante na inferência estatística e em combinações de amostras.

Desvio Padrão (S)

O desvio-padrão é a medida mais usada na comparação de diferenças entre conjuntos de dados, por ter grande precisão. O desvio padrão determina a dispersão dos valores em relação à média e é calculado por meio da raiz quadrada da variância.

$$S=\sqrt{s^2}$$

Coeficiente de Variação (CV)

Trata-se de uma medida relativa de dispersão útil para a comparação em termos relativos do grau de concentração. O coeficiente de variação é a relação entre o desvio padrão (S) e a média $\bar{x}$.

$$CV = \frac{S}{\bar{x}}$$

Diz- se que uma distribuição tem:

Baixa dispersão: CV ≤ 15%

Média dispersão: 15%< CV<30%

Alta dispersão: CV ≥ 30%

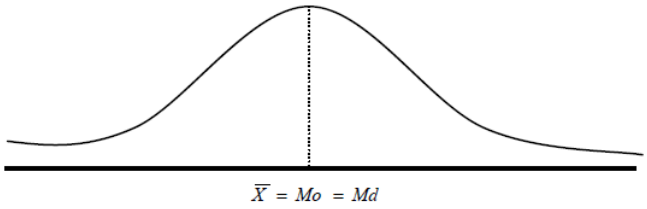
Assimetria

Estas medidas referem-se à forma da curva de uma distribuição de frequência, mais especificamente do polígono de frequência ou do histograma.

Denomina-se assimetria o grau de afastamento de uma distribuição da unidade de simetria.

• Em uma distribuição simétrica, tem-se igualdade dos valores da média, mediana e moda.

Simetria

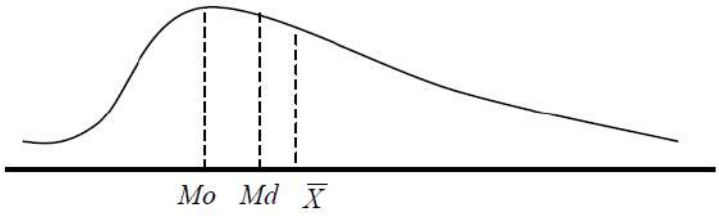


Toda distribuição deformada é sempre assimétrica. Entretanto, a assimetria pode dar-se na cauda esquerda ou na direita da curva de frequências.

• Em uma distribuição assimétrica positiva, ou assimetria à direita, tem-se:

Assimetria à direita (ou positiva)

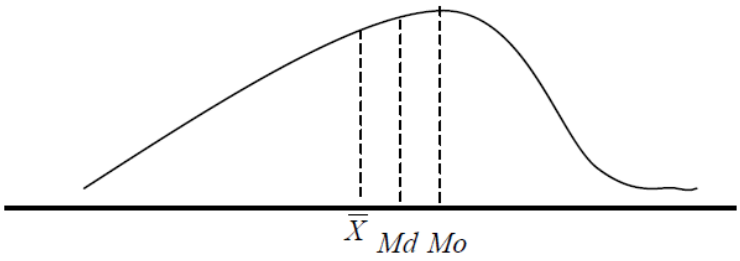
$Mo < Md < \overline{X}$



• Em uma distribuição assimétrica negativa, ou assimetria à esquerda, predominam valores inferiores à Moda.

Assimetria à esquerda (ou negativa)

$$\overline{X} < Md < Mo$$



Existem várias fórmulas para o cálculo do coeficiente de assimetria. As mais utilizadas são:

• 1º Coeficiente de Pearson

$$AS = \frac{\overline{x} - Mo}{S}$$

M_o: valor modal (moda)
S : Desvio padrão
$\overline{X}$: Média

• 2º Coeficiente de Pearson

$$AS = \frac{Q_1 + Q_3 - 2Md}{Q_3 - Q_1}$$

Q_1 : valor do 1º Quartil
Q_3 : valor do 3º Quartil
M_d : valor da Mediana

Quando

AS = 0, diz-se que a distribuição é simétrica.

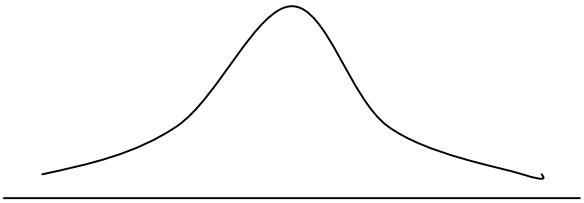
AS > 0, diz-se que a distribuição é assimétrica positiva (à direita)

AS > 0, diz-se que a distribuição é assimétrica positiva (à direita)

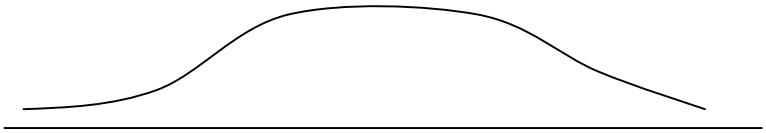
Curtose

Curtose é o grau de achatamento (ou afilamento) de uma distribuição em comparação com uma distribuição padrão (chamada curva normal). De acordo com o grau de curtose, classificamos três tipos de curvas de frequência:

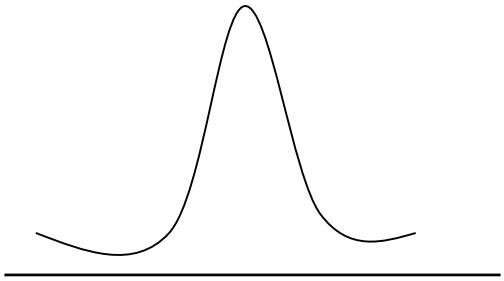
- Mesocúrtica: é uma curva básica de referência chamada curva padrão ou curva normal;



- Platicúrtica: é uma curva mais achatada (ou mais aberta) que a curva normal;



- Leptocúrtica: é uma curva mais afilada que a curva normal;



Para medir o grau de curtose utiliza-se o coeficiente:

$$K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$

Q_1 : valor do 1º Quartil
Q_3 : valor do 3º Quartil
P_{10} : valor do percentil 10
P_{90} : valor do percentil 90

- Se K= 0,263, diz-se que a curva correspondente à distribuição de frequência é mesocúrtica.

- Se K > 0,263, diz-se que a curva correspondente à distribuição de frequência é platicúrtica.
- Se K < 0,263, diz-se que a curva correspondente à distribuição de frequência é leptocúrtica.

Probabilidade

Vamos ver que a probabilidade expressa por meio de valores numéricos as possibilidades da ocorrência dos resultados de um fenômeno.

Experimento aleatório, espaço amostral e eventos

Introdução

Todas as vezes que se estudam fenômenos de observação, cumpre-se distinguir o próprio fenômeno e o modelo matemático (determinístico ou probabilístico) que melhor o explique.

Os fenômenos estudados pela estatística são fenômenos cujo resultado, mesmo em condições normais de experimentação, varia de uma observação para outra, dificultando dessa maneira a previsão de um resultado.

A observação de um fenômeno casual é recurso poderoso para se entender a variabilidade do mesmo. Entretanto, com suposições adequadas e sem observar diretamente o fenômeno, podemos criar um modelo teórico que reproduza de forma bastante satisfatória a distribuição das frequências quando o fenômeno é observado diretamente. Tais modelos são os chamados modelos de probabilidades.

Os fenômenos determinísticos conduzem sempre a um mesmo resultado quando as condições iniciais são as mesmas. Ex: tempo de queda livre de um corpo. Mantidas as mesmas condições, as variações obtidas para o valor do tempo de queda livre de um corpo são extremamente pequenas (em alguns casos, desprezíveis).

Os fenômenos aleatórios podem conduzir a diferentes resultados e mesmo quando as condições iniciais são as mesmas, existe a imprevisibilidade do resultado. Ex: lançamento de um dado.

Podemos considerar os experimentos aleatórios como fenômenos produzidos pelo homem.

- Lançamento de uma moeda honesta;
- Lançamento de um dado;
- Lançamento de duas moedas;
  - Retirada de uma carta de um baralho completo, de 52 cartas;
  - Determinação da vida útil de um componente eletrônico.
  - A análise desses experimentos revela que:
  - Cada experimento poderá ser repetido indefinidamente sob as mesmas condições;
  - Não se conhece em particular valor o do experimento “a priori”, porém pode-se descrever todos os possíveis resultados – as possibilidades;
  - Quando o experimento for repetido um grande número de vezes, surgirá uma regularidade.

Para a explicação desses fenômenos (fenômenos aleatórios), adota-se um modelo matemático probabilístico.

Para melhor entendimento é interessante relembrar alguns conceitos básicos no estudo das probabilidades tais como:

Espaço Amostral

Um dos conceitos matemáticos fundamentais utilizados no estudo das probabilidades é o de conjunto. Um conjunto é uma coleção de objetos ou itens que possuem característica(s) comum(ns). É importante definir cuidadosamente o que constitui o conjunto em que estamos interessados, a fim de podermos decidir se determinado elemento é ou não membro do conjunto.

Conjunto é uma coleção bem definida de objetos ou itens.

A probabilidade só tem sentido no contexto de um espaço amostral, que é o conjunto de todos os resultados possíveis de um "experimento". O termo "experimento" sugere a incerteza do resultado antes de fazermos as observações. Os resultados de um experimento (ex: a ocorrência de um raio, uma viagem etc.) chamamse eventos.

Evento Aleatório (E)

É qualquer subconjunto de um espaço amostral. É também o resultado obtido de cada experimento aleatório, que não é previsível.

Espaço Amostral (S)

Espaço amostral de um experimento aleatório é o conjunto de todos os possíveis resultados desse experimento.

Exemplos de Espaços Amostrais:

- S = { c, r } (é composto de 2 eventos)
- S = { 1, 2, 3, 4, 5, 6 } (é composto de 6 eventos)
- S = { (c, r), (c, c), (r, c), (r, r)} (é composto de 8 eventos)

Genericamente, se o nº de pontos (elementos do espaço amostral) amostrais de um espaço amostral finito é n, então o número de eventos é dado por 2^n.

Exemplo: No lançamento de 5 moedas, o número de pontos amostrais (resultados possíveis) é $2^5 = 32$. Portanto, S= 32.

O complemento de um evento é constituído de todos os resultados no espaço amostral que não façam parte do evento.

Os eventos são mutuamente excludentes, quando não têm elemento em comum.

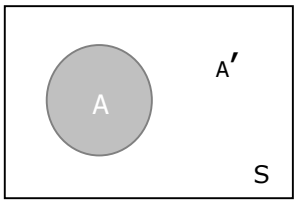
Ou se não podem ocorrer simultaneamente.

Exemplo:

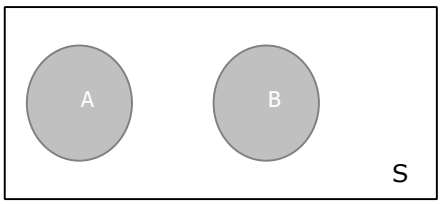
Na extração de uma só carta, os eventos "a carta é de copas" e a "carta é de ouros" são mutuamente excludentes, porque uma carta não pode ser ao mesmo tempo de copas e de ouros.

Já os eventos "a carta é de copas" e "a carta é uma figura" não são mutuamente excludentes, porque algumas cartas de copas são também figuras.

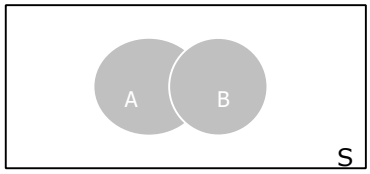
Muitas vezes, é útil representar graficamente um espaço amostral, porque isso torna mais fácil visualizar lhe os elementos.



Os eventos A e A' são complementares.



Os eventos A e B são mutuamente excludentes porque não se interceptam.



Os eventos A e B não são mutuamente excludentes, pois têm alguns elementos em comum.

Operações com Eventos Aleatórios

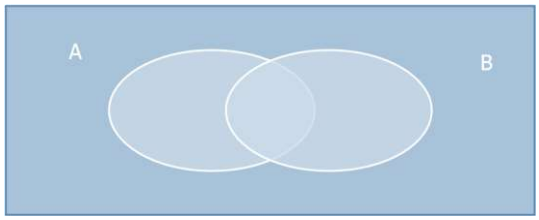
Consideremos um espaço amostral finito:

$$S = \{\in_1, \in_2, \in_3 \ldots\ \in_n\}$$

Sejam A e B dois eventos de S, as seguintes operações são definidas:

a) Reunião – A ∪ B – O evento reunião é formado pelos pontos amostrais que pertencem a pelo menos um dos eventos.

$$A \cup B = \{\in_1 (\in) S \,/\, \in_1 \in A \text{ ou } \in_1 \in B\}$$

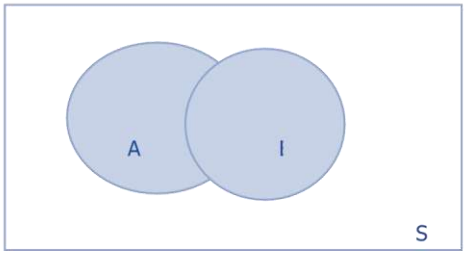


É o evento que ocorre se A ocorre ou B ocorre, ou ambos ocorrem

b) Interseção – A ∩ B – O evento interseção é formado pelos pontos amostrais que pertencem simultaneamente aos eventos A e B.

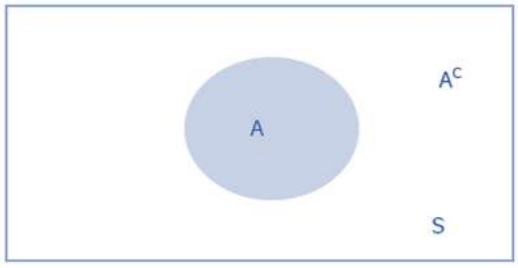
$$A \cap B = \{\epsilon_i \in S | \epsilon_i \in A \; e \; \epsilon_i \in B\}$$

É o evento que ocorre se A e B ocorrem



Obs: Se A ∩ B = ∅, A e B são eventos mutuamente exclusivos

c) Complementação → S – A = A^c → É o evento que ocorre se A não ocorre.



$$A^c = \{\epsilon_i \in S | \epsilon_i \notin A\}$$

Probabilidade: Definição Clássica; Probabilidade e frequência relativa

Dado um experimento aleatório, E e S o espaço amostral, probabilidade de um evento A – P(A) - é uma função definida em S que associa a cada evento um número real, satisfazendo os seguintes axiomas:

- 0 < P(A) <1
- P(S) = 1
- Se A e B forem eventos mutuamente exclusivos, (A ∩ B)=Ø), então:

P (A ∪ B)=P(A)+P(B)

Chamamos de probabilidade de um evento A (A ⊂ S) o número real P(A), tal que:

$$P(A) = \frac{\text{Número de Casos Favoráveis (A)}}{\text{Número Total de Casos}} = \frac{NCF(A)}{NTC}$$

Tipos de eventos

Chamamos de evento qualquer subconjunto do espaço amostral S de um experimento aleatório.

Assim, qualquer que seja E, se E ⊂ S (E está contido em S), então E é um evento de S.

Evento Certo – é aquele que ocorre em qualquer realização do experimento aleatório.

Se E = S, E é chamado evento certo.

Evento Elementar – é aquele formado por um único elemento do espaço amostral.

Se E ⊂ S e E é um conjunto unitário, E é chamado evento elementar.

Evento Impossível - é aquele que não ocorre em nenhuma realização de um experimento aleatório. Se E = ∅, E é chamado evento impossível.

Evento Complementar – seja um evento A qualquer, o evento A' (chamado de complementar de A), tal que A'=S-A, ou seja, é um outro conjunto formado pelos elementos que pertencem a S e não pertencem a A.

Eventos Equiprováveis - Quando se associa a cada ponto amostral a mesma probabilidade, o espaço amostral chama-se equiprovável ou uniforme. Os eventos E_i,i=1,2,3.....n são equiprováveis quando $P(E_i)=P(E_2)=P(E_n)=P$, isto é, quando todos têm a mesma probabilidade de ocorrer: P=1/n.

Se os n pontos amostrais (eventos) são equiprováveis, a probabilidade de cada um dos pontos amostrais é 1/n.

$$P(A) = r\left(\frac{1}{n}\right) \therefore P(A) = \frac{r}{n}$$

Exemplo:

Retira-se uma carta de um baralho completo de 52 cartas. Qual a probabilidade de sair um rei ou uma carta de espadas?

Solução:

$$P(A \cup B) = P(A)+P(B)-P(A \cap B)$$

Eventos mutuamente exclusivos

Dois eventos A e B são denominados mutuamente exclusivos se eles não puderem ocorrer simultaneamente, isto é, $A \cap B = \emptyset$.

Evento impossível

Exemplo:

E = Jogar um dado e observar o resultado. Sejam os eventos A= ocorrer nº par e B= ocorrer nº ímpar.

$$S=\{1,2,3,4,5,6\}$$

Então, A e B são mutuamente exclusivos, pois a ocorrência de um número par e ímpar não pode ser verificada em decorrência da mesma experiência.

Axiomas de Probabilidade

1 – Se Ø é o conjunto vazio (evento impossível), então P(Ø)=0.

2 – Teorema do evento complementar: Se A^c é o complemento do evento A, então:

$$P(Ac) = 1 - P(A)$$

3 – Teorema da soma: Se A e B são dois eventos quaisquer, então:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Se tomássemos apenas P(A ∪ B) = P (A) + P (B), estaríamos, considerando duas vezes a probabilidade de interseção.

Teoria da Contagem

Dados dois eventos, o primeiro dos quais pode ocorrer de m maneiras distintas e o segundo pode ocorrer de n maneiras distintas, então os dois eventos conjuntamente podem ocorrer de m.n maneiras distintas. O cálculo da probabilidade de um evento reduz-se a um problema de contagem. Assim é que a Análise Combinatória tem fundamental importância para se contar o nº de casos favoráveis e o total de casos. Para problemas simples ou com poucos elementos, pode-se contr o númro de resultados de forma direta, sem necessidade de recorrer às fórmulas matemáticas da análise combinatória. Para problemas menos simples, recorre-se ás combinações e arranjos para determinar o número de casos.

- **Combinação**
  - O Número de combinações de r elementos combinados p a p sendo p<r e calculado por:

$$C_{r,p} = \binom{r}{p} = \frac{r!}{p!(r-p)!}$$

Notação: O Símbolo **fatorial !** denota o produto dos inteiros positivos em ordem decrescente. Por exemplo, 6! = 6.5.4.3.2.1= 720. Por definição , 0! = 1.

Exemplo: Quantas comissões de três pessoas podem ser formadas com um grupo de dez pessoas?

$$C_{10,3} = \binom{10}{3} = \frac{10.9.8.7!}{3.2.7!} = 120$$

Podemos ter 120 comissões diferentes compostos com 3 pessoas.

- **Arranjos**

o O número de arranjos de r elementos é calculado por:

$$A_{r,p} = \frac{r!}{(r-p)!}$$

Exemplo: Considerando um grupo de dez pessoas, quantas chapas diferentes podemos ter para uma eleição de presidente, tesoureiro e secretário?

$$A_{10,3} = \frac{10!}{(10-3)!} = \frac{3628800}{7!} = \frac{3628800}{5040} = 720$$

Podemos ter 720 chapas diferentes**.**

OBS: Quando queremos selecionar r elementos de um conjunto de n elementos distintos sem levar em conta a ordem, estamos considerando combinações, quando contamos separadamente ordenações diferentes dos mesmos elementos temos arranjos.

Probabilidade Condicional e Independência de Eventos

Probabilidade Condicional

O evento em que ambos, A e B, ocorrem é chamado A interseção B; portanto, a probabilidade do evento A ocorrer, dado que B ocorreu, é de:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Isso significa que a probabilidade de A ocorrer, dado que B ocorreu, é igual à probabilidade de ocorrência simultânea de A e B dividida pela probabilidade de ocorrência de B. (Note-se que essa definição não se aplica quando P(B)=0, porque então estaríamos dividindo por zero).

Exemplo:

Dois dados são lançados. Consideremos os eventos

$A=\{(x_1,x_2) / x_1+x_2=10\}$ e $B=\{(x_1,x_2) / x_1 > x_2\}$,

onde x_1 é o resultado do dado 1 e x_2 é o resultado do dado 2. Avaliar P(A); P(B); P(A/B) e P(B/A).

Solução:

P(A)= 3/36 =1/12

P(B) =15/36 =5/12

P(A/B)=1/15

P(B/A)=1/3

Teorema do Produto

"A probabilidade da ocorrência simultânea de dois eventos, A e B, do mesmo espaço amostral, é igual ao produto da probabilidade de um deles pela probabilidade condicional do outro, dado o primeiro".

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \qquad P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Independência Estatística

Um evento A é considerado independente de um outro evento B se a probabilidade de A é igual à probabilidade condicional de A dado B.

Exemplo: Em um lote de 12 peças, 4 são defeituosas, 2 peças são retiradas uma após a outra, sem reposição. Qual a probabilidade de que ambas sejam boas?

A={a 1ª peça é boa}

B={a 2ª peça é boa}

P(A ∩ B)= P(A). P(B/A)= 8/12.7/11=14/33

Isto é, se P(A)=P(A/B) «É evidente que, se A é independente de B, B é independente de A; P(B)=P(B/A). Se A e B são independentes, então temos que

P(A ∩ B)=P(A).P(B)

Regra de Bayes

Sejam A_1, A_2, A_3,.....A_n, n eventos mutuamente exclusivos tais que $A_1 \cup A_2 \cup A_n = S$.

Sejam $P(A_i)$ as probabilidades conhecidas dos vários eventos e B um evento qualquer de S, tal que são conhecidas todas as probabilidades condicionais $P(B/A_i)$.

$$P(A_i / B) = \frac{P(A_i).P(B / A_i)}{P(A_1).P(B / A_1) + P(A_2).P(B / A_2) + ... + P(A_n).P(B / A_n)}$$

OBS: O Teorema de Bayes é também chamado de Teorema da Probabilidade a Posteriori. Ele relaciona uma das parcelas da probabilidade total com a própria probabilidade total.

É uma generalização da probabilidade condicional ao caso de mais de dois eventos.

Exemplos:

1. Sendo P(A) = 1/3, P(B)= ¾ e P(A ∪ B)=11/12, calcular P(A/B).

   Solução:

   Como $P(A/B) = \frac{P(A \cap B)}{P(B)}$, devemos calcular P(A ∩ B).

   Como P(A ∪ B) = P(A) + P(B) - P(A ∩ B),

   temos: 11/12= 1/3 + ¾ - P(A ∩ B)

∴ P(A ∩ B)= 2/12=1/6

logo, P(A/B) = $\frac{1/6}{3/4} = \frac{2}{9}$

2. Em certo colégio, 5% dos homens e 2% das mulheres têm mais do que 1,80 m de altura. Por outro lado, 60% dos estudantes são homens. Se um estudante é selecionado aleatoriamente e tem mais de 1,80m de altura, qual a probabilidade de que o estudante seja mulher?
   Solução:
   Temos que :

   P(Ma/H) = 0,05 (Probabilidade de Homem ter mais de 1,80 m)

   P(Ma/M) = 0,02 (Probabilidade de Mulher ter mais de 1,80 m)

   P(H) = 0,6 (Probabilidade de ser homem)

   P(M) = 0,4 (Probabilidade de ser mulher)

   P(M/Ma)= ? (Probabilidade de ser mulher dado que tem mais que 1,80 m)

   Utilizando a Regra de Bayes temos:

$$P(M/Ma) = \frac{P(M)*P(Ma/M)}{P(M)*P(Ma/M)+P(H)*P(Ma/H)} = \frac{0,4*0,02}{(0,4*0,02)+(0,6*0,05)}$$

$$P(M/Ma) = \frac{0,008}{0,038} = 0,21 = 21\%$$

3. Três máquinas, A, B e C produzem respectivamente 40%, 50% e 10% do total de peças de uma fábrica. As porcentagens de peças defeituosas nas respectivas máquinas são 3%, 5% e 2%. Uma peça é sorteada ao acaso e verifica-se que é defeituosa. Qual a probabilidade de que a peça tenha vindo da máquina B? E da máquina A?

   Solução

Temos que :

P(A) = 0,4

P(B) = 0,5

P(C)= 0,10

P(D/A) = 0,03

P(D/B) = 0,05

P(D/C) = 0,02

P(B/D) = ?

Utilizando a Regra de Bayes temos:

$$P(B/D) = \frac{P(B) * P(D/B)}{P(A) * P(D/A) + P(B) * P(D/B) + P(C) * P(D/C)}$$

$$P(B/D) = \frac{0,5 * 0,05}{0,4 * 0,03 + 0,5 * 0,05 + 0,1 * 0,02}$$

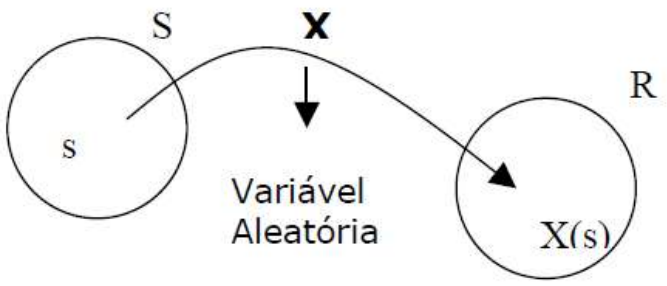
$$P(B/D) = \frac{0,025}{0,039} = 0,641 = 64,1\%$$

Variáveis Aleatórias

Muitos experimentos aleatórios produzem resultados não-numéricos. Antes de analisá-los, é conveniente transformar seus resultados em números, o que é feito através da variável aleatória, que é uma regra de associação de um valor numérico a cada ponto do espaço amostral. Portanto, variáveis aleatórias são variáveis numéricas às quais iremos associar modelos probabilísticos. Veremos que uma variável aleatória tem um número para cada resultado de um experimento e que uma distribuição de probabilidades associa uma probabilidade a cada resultado numérico de um experimento.

Conceito de variável aleatória

Sejam E um experimento e S o espaço associado ao experimento. Uma função X, que associe a cada elemento s ∈ S um número real X(s), é denominada variável aleatória.



Exemplo:

E: lançamento de duas moedas;

X: nº de caras obtidas nas duas moedas;

S={(C,C), (C,R), (R,C),(R,R)}

X=0 → corresponde ao evento (r,r) com probabilidade ¼

X= 1→ corresponde ao evento (r,c), (c,r) com probabilidade 2/4

X= 2→ corresponde ao evento (c,c) com probabilidade ¼.

Empregamos a termo variável aleatória para descrever o valor que corresponde ao resultado de determinado experimento.

As variáveis aleatórias também podem ser discretas ou continuas e temos as seguintes definições:

Variáveis Aleatórias Discretas – Admite um número finito de valores ou tem uma quantidade enumerável de valores.

Variáveis Aleatórias Continuas – pode tomar um número infinito de valores, e esses valores podem ser associados a mensurações em uma escala contínua.

Distribuição de probabilidade

Uma vez definida a variável aleatória, existe interesse no cálculo dos valores das probabilidades correspondentes.

O conjunto das variáveis e das probabilidades correspondentes é denominado distribuição de probabilidades, isto é:

$$\{(x_i, p(x_i), I=1,2,\ldots n\}$$

Função de densidade de probabilidade

É a função que associa a cada valor assumido pela variável aleatória a probabilidade do evento correspondente, isto é:

$$P(X=x_i)= P(A_i), i=1,2,\ldots,n$$

Esperança matemática, variância e desvio padrão: propriedades

Existem características numéricas que são muito importantes em uma distribuição de probabilidades de uma variável aleatória discreta. São os parâmetros das distribuições, a saber:

Esperança matemática (ou simplesmente média) - E (x) – é um número real, é também uma média aritmética;

Variância - VAR (x) – é a medida que dá o grau de dispersão (ou de concentração) de probabilidade em torno da média. O fato de conhecermos a média de uma distribuição de probabilidades já nos ajuda bastante, porém, precisamos de uma medida que nos dê o grau de dispersão de probabilidade em torno dessa média.

Desvio Padrão – DP(X) – é a raiz quadrada da variância.

Distribuições discretas: Bernoulli, Binomial e Poisson

Uma função X, definida sobre o espaço amostral e assumindo valores num conjunto enumerável de pontos do conjunto real, é dita uma variável aleatória discreta.

Uma variável aleatória X do tipo discreto estará bem caracterizada se indicarmos os possíveis valores x_1, x_2, …, x_k que ela pode assumir e as respectivas probabilidades $p(x_1)$, $p(x_2)$, …, $p(x_k)$, ou seja, se conhecermos a sua função de probabilidade (x; p(x)), onde

$$p(x) = P(X = x)$$

Dada a v.a. discreta X, assumindo os valores x_1, x_2, …, x_k , chamamos esperança matemática de X ao valor:

$$E(X) = \sum x \cdot p(x)$$

Chamamos variância de X ao valor:

$$Var(X) = E(X^2) - [E(X)]^2$$

onde

$$E(X^2) = \sum x^2 \cdot p(x)$$

E de desvio padrão de X a

$$DP(X) = \sqrt{Var(X)}$$

Distribuições Discretas de Probabilidade

Distribuição de Bernoulli

Seja um experimento aleatório E realizado repetidas vezes, sempre nas mesmas condições, de tal forma que o resultado pode ser um Sucesso (s) **(**se acontecer o evento que nos interessa) ou um **Fracasso** (f) (se o evento não se realizar).

Seja X a variável aleatória: Sucesso ou Fracasso

$$X \to x_1 = 1 \text{ (sucesso)} \quad \text{ou} \quad x_2 = 0 \text{ (fracasso)}$$
$$P(X) \to p(x_1) = p \qquad p(x_2) = 1 - p = q$$
$$P(x = 0) = q \quad \text{e} \quad P(x = 1) = p$$

Essas condições caracterizam um conjunto de Provas de Bernoulli ou um experimento de Bernoulli, e sua função probabilidade é dada por:

$$P(X = x) = p^{x}.q^{1-x}$$

Principais características

- Média:

$$E(X) = \sum_{i=0}^{1} x_i P(x_i) = 0.q + 1.p = p$$

- Variância:

$$Var(X) = E(X^2) - [E(X)]^2$$
$$E(X^2) = \sum_{i=0}^{1} x_i^2 P(x_i) = 0^2 q + 1^2 p = p$$
$$Var(X) = p - p^2 = p(1-p) = pq$$

Distribuição Binomial

Uma variável aleatória tem distribuição binomial quando o experimento ao qual está relacionada apresenta apenas dois resultados (sucesso ou fracasso). Este modelo fundamenta-se nas seguintes hipóteses:

- n provas independentes e do mesmo tipo são realizadas;
- cada prova admite dois resultados – Sucesso ou Fracasso;
- a probabilidade de sucesso em cada prova é p e de fracasso 1- p = q

Define-se a Variável X que conta o número de sucesso nas n realizações do experimento.

(X pode assumir os valores 0, 1, 2, 3, ..., n.)

Fazendo sucesso corresponder a 1 e fracasso, a 0, temos:

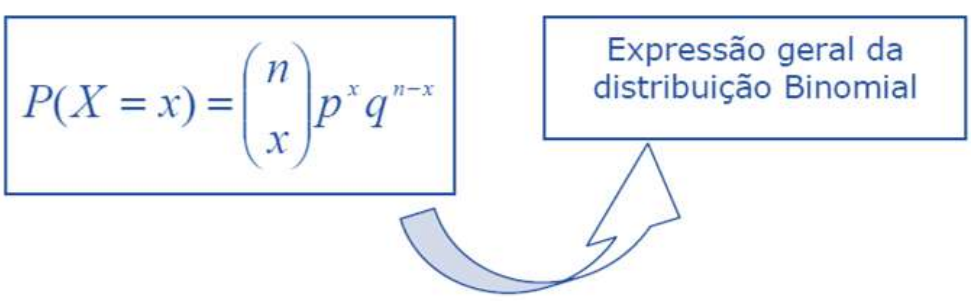
- Para X = 0 , uma sequência de n zeros : 000000....000.

$$P(X = 0) = q.q.q.q.q.....q = q^n$$

- Para X = 1, uma sequência do tipo: 1000....0; 01000....0; 001000...0; serão n sequências, cada uma com um único sucesso e n-1 fracassos:

$$P(X = 1) = n.p.q^{n-1}$$

- Para X= x, tem-se x sucessos e (n-x) fracassos, correspondendo às sequências com x algarismos 1 e n-x zeros. Cada sequência terá probabilidade $p^x q^{n-x}$ e como há $\frac{n}{x}$ sequências distintas, tem-se:



Principais características:

- Média: E(X) = n.p
- Variância: Var(X) = n.p.q

Exemplo: Uma moeda não viciada é lançada 8 vezes. Encontre a probabilidade de:

a) dar 5 caras;
b) pelo menos uma cara;
c) no máximo 2 caras.

Solução:

a) x = sair cara, p=0,5 (probabilidade do sucesso de X), q= 0,5 (probabilidade do fracasso de X0, n = 8 (número de repetições do evento).

$$P(X=5)=\frac{8!}{5!(8-5)!}.0{,}5^5.0{,}5^{8-5}=\frac{40320}{120.6}.0{,}5^8=0{,}21875=21{,}88\%$$

b) $P(x\geq 1)= 1 – \{ P(X=0)\} = 1-\{ q^n\} = 1- 0{,}5^8 = 0{,}9960 = 99{,}6\%$

c) $P(X\leq 2) = P(X=0)+ P(x=1) + P(X=2)$, utilizando as fórmulas dos itens anteriores calcula-se as probabilidades.

Distribuição de Poisson

Consideremos a probabilidade de ocorrência de sucessos em um determinado intervalo. A probabilidade da ocorrência de um sucesso no intervalo é proporcional ao intervalo. A probabilidade de mais de um sucesso nesse intervalo é bastante pequena em relação à probabilidade de um sucesso.

Seja X o número de sucessos no intervalo; temos, então:

$$P(X = k) = \frac{e^{-\lambda} . \lambda^{k}}{k!}$$

onde λ é a média.

A variável X assim definida tem distribuição de Poisson.

A distribuição de Poisson é muito usada na distribuição do número de:

- Carros que passam por um cruzamento por minuto, durante uma certa hora do dia;
- Erros tipográficos por página, em um material impresso;
- Defeitos por unidade (m^3, m^2, m, etc.,,) por peça fabricada;
- Colônias de bactérias numa dada cultura por 0,01mm^2, numa plaqueta de microscópio;
- Mortes por ataque de coração por ano, numa cidade;
- Problemas de filas de espera em geral, e outros.

Principais características:

- Média**:** E(X) = λ
- Variância: Var(X) = λ

OBS: Muitas vezes, no uso da binomial, acontece que n é muito grande e p é muito pequeno. Podemos, então, fazer uma aproximação de binomial pela distribuição de Poisson, da seguinte forma:

λ= np

Exemplo: Em média, são feitas 2 chamadas por hora para certo telefone. Calcular a probabilidade de se receber no máximo 3 chamadas em 2 horas e a probabilidade de nenhuma chamada em 90 minutos.

Solução:

λ = 2 por hora, para 2 horas λ= 4

P(X≤3)= P(X=0)+P(X=1)+P(X=2)+P(X=3)∴

$$P(X\leq 3) = \frac{e^{-\lambda}.\lambda^{0}}{0!} + \frac{e^{-\lambda}.\lambda^{1}}{1!} + \frac{e^{-\lambda}.\lambda^{2}}{2!} + \frac{e^{-\lambda}.\lambda^{3}}{3!} \therefore$$

$$P(X\leq 3) = \frac{e^{-4}.4^{0}}{1} + \frac{e^{-4}.4^{1}}{1} + \frac{e^{-4}.4^{2}}{2} + \frac{e^{-4}.4^{3}}{6} \therefore$$

$$P(X\leq 3) = e^{-4} + 4e^{-4} + 8e^{-4} + 10{,}666e^{-4} = 23{,}66e^{4} = 0{,}4334 = 43{,}34\%$$

Agora para 90 minutos λ=3 (1 hora (60 minutos) λ=2)

$$P(X=0) = \frac{e^{-3}.3^{0}}{1} = e^{-3} = 0{,}0498 = 4{,}98\ \%$$

Distribuição contínua: Normal - propriedades, distribuição normal padrão, a Normal como aproximação da Binomial

Uma variável aleatória, cujos valores são expressos em uma escala contínua, é chamada de variável aleatória contínua.

Podemos construir modelos teóricos para v.a.'s contínuas, escolhendo adequadamente a função de densidade de probabilidade (f.d.p.**)**, que é uma função indicadora da probabilidade nos possíveis valores de X.

Assim, a área sob a f.d.p. entre dois pontos a e b nos dá a probabilidade da variável assumir valores entre a e b, conforme ilustrado na figura 1, apresentada a seguir.

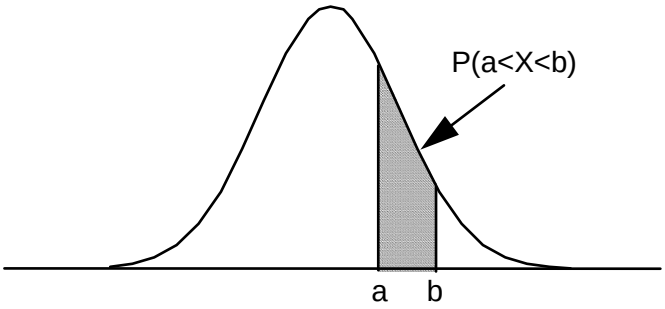


Figura 1 – Probabilidade como área sob a curva entre dois pontos

Portanto, podemos escrever:

$$P(a < X < b) = \int_a^b f(x)dx$$

Da relação entre a probabilidade e a área sob a função, a inclusão, ou não, dos extremos a e b na expressão acima não afetará os resultados. Assim, iremos admitir

$$\mathbf{P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)}$$

Teoricamente, qualquer função f(x) que seja não negativa e cuja área total sob a curva seja igual à unidade, isto é,

$$\int f(x)dx = 1$$

caracterizará uma v.a. contínua.

Dada a v.a. contínua X, assumindo os valores no intervalo entre a e b, chamamos valor médio ou esperança matemática de X ao valor

$$E(X) = \int_a^b x \cdot f(x)dx$$

Chamamos variância de X ao valor

$$Var(X) = E(X^2) - [E(X)]^2$$

onde

$$E(X^2) = \int_a^b x^2 \cdot f(x)dx$$

e de desvio padrão de X a

$$\mathbf{DP(X) = \sqrt{Var(X)}}$$

Se X é uma v.a. contínua com f.d.p. f(x), definimos a sua função de distribuição acumulada F(x) como:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Distribuição Normal

A distribuição normal é a mais importante das distribuições de probabilidades. Conhecida como a "curva em forma de sino", a distribuição normal tem sua origem associada aos erros de mensuração. É sabido que quando se efetuam repetidas mensurações de determinada grandeza com um aparelho equilibrado, não se chega ao mesmo resultado todas as vezes; obtém-se, ao contrário, um conjunto de valores que oscilam, de modo aproximadamente simétrico, em torno do verdadeiro valor. Construindo-se o histograma desses valores, obtém-se uma figura com forma aproximadamente simétrica. Gauss deduziu matematicamente a distribuição normal como distribuição de probabilidade dos erros de observação, denominando-a então "lei normal dos erros".

Supunha-se inicialmente que todos os fenômenos da vida real devessem ajustar-se a uma curva em forma de sino; em caso contrário, suspeitava-se de alguma anormalidade no processo de coleta de dados. Daí a designação de curva normal.

A observação cuidadosa subsequente mostrou, entretanto, que essa pretensa universalidade da curva, ou distribuição normal, não correspondia à realidade. De fato, não são poucos os exemplos de fenômenos da vida real representados por distribuições não normais, curvas assimétricas, por exemplo. Mesmo assim, a distribuição normal desempenha papel preponderante na estatística, e os processos de inferência nela baseados têm larga aplicação.

A distribuição normal tem sua função de densidade de probabilidade dada por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad -\infty < x < \infty$$

Como pode-se observar através da equação acima, a distribuição normal inclui os parâmetros μ e σ, os quais possuem os seguintes significados:

μ : posição central da distribuição (média, μ_x)

σ : dispersão da distribuição (desvio padrão, σ_x)

Se uma variável aleatória X tem distribuição normal com média μ e variância σ^2, escrevemos: $X \sim N(\mu, \sigma^2)$.

A figura 2 ilustra uma curva normal típica, com seus parâmetros descritos graficamente.

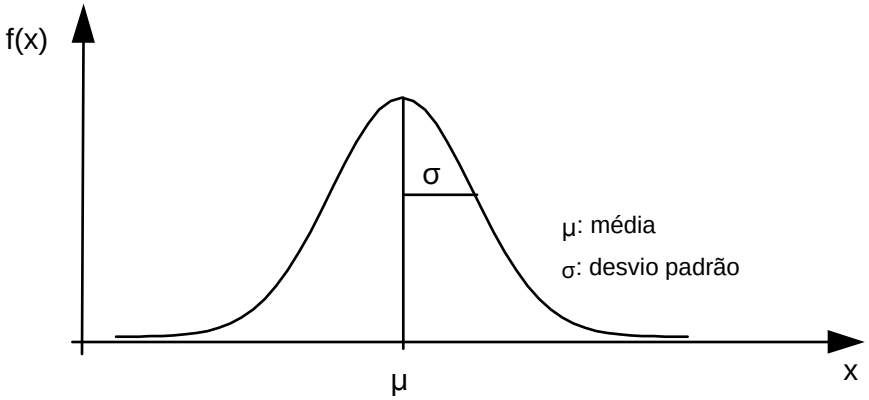


Figura 2 - Curva normal típica

Propriedades da distribuição normal

Para uma mesma média μ e diferentes desvios padrão σ, a distribuição que tem maior desvio padrão se apresenta mais achatada, acusando maior dispersão em torno da média. A que tem menor desvio padrão apresenta "pico" mais acentuado e maior concentração em torno da média. A figura 3 compara três curvas normais, com a mesma média, porém, com desvios padrão diferentes. A curva A se apresenta mais dispersa que a curva B, que por sua vez se apresenta mais dispersa que a curva C. Nesse caso, $\sigma_A > \sigma_B > \sigma_C$.

Distribuições normais com o mesmo desvio padrão e médias diferentes possuem a mesma dispersão, mas diferem quanto à localização. Quanto maior a média, mais à direita está a curva. A figura 4 ilustra o fato, onde a curva A possui média maior que a curva B ($\mu_A > \mu_B$).

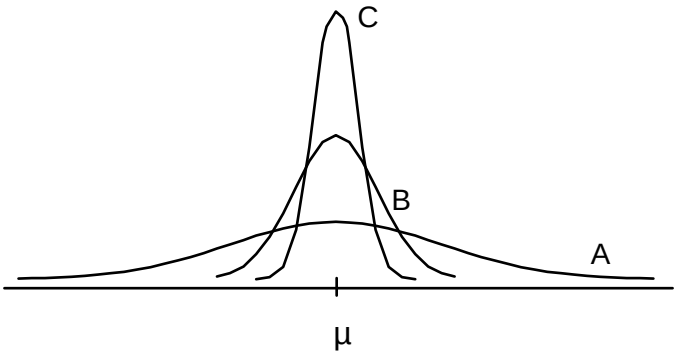


Figura 3 - Distribuições normais com mesma média e desvios padrão diferentes

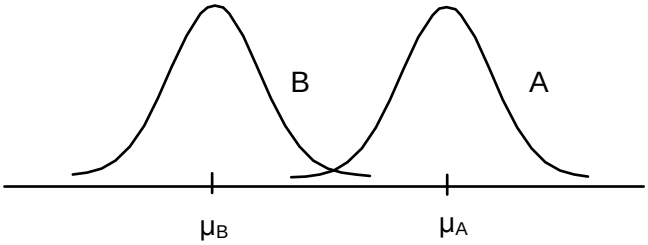


Figura 4 - Distribuições normais com mesmo desvio padrão e médias diferentes

Como descrito anteriormente, a probabilidade de uma variável assumir valores entre a e b é igual à área sob a curva entre esses dois pontos. A determinação dessas probabilidades é realizada matematicamente através da integração da função de densidade de probabilidade entre os pontos a e b de interesse. No caso da normal, a integral não pode ser calculada exatamente e a probabilidade entre dois pontos só pode ser obtida de forma aproximada, por métodos numéricos. Essa tarefa é facilitada através do uso da distribuição normal padrão definida a seguir.

No caso da distribuição normal, algumas dessas áreas - com os pontos a e b, função da média μ e do desvio padrão σ - são bastante difundidas e estão representadas na figura 5:

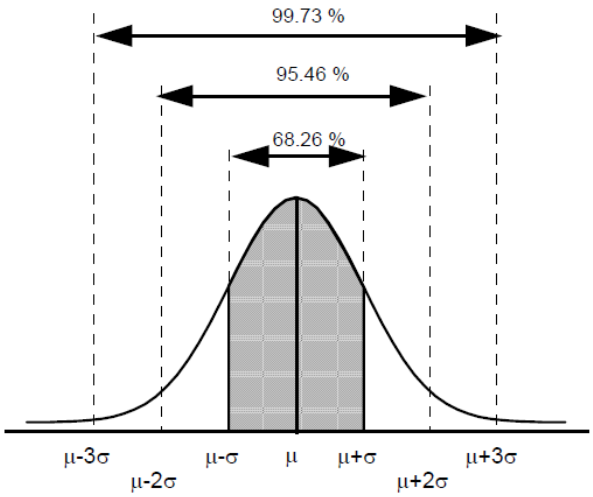


Figura 5 - Probabilidades da distribuição normal

Portanto, 68,26% dos valores populacionais caem entre os limites definidos como média mais ou menos um desvio padrão ($\mu \pm 1^{\sigma}$); 95,46% dos valores caem entre média mais ou menos dois desvios padrão ($\mu \pm 2\sigma$); e 99,73% dos valores caem entre média mais ou menos três desvios padrão ($\mu \pm 3\sigma$).

A distribuição normal padrão

A distribuição normal particular com média 0 e desvio padrão 1 é chamada de distribuição normal padrão e costuma ser denotada por Z.

Se $X \sim N(\mu, \sigma^2)$, então, a variável aleatória definida por

$$Z = \frac{X - \mu}{\sigma}$$

terá uma distribuição N(0,1). Essa transformação é ilustrada pela figura 6:

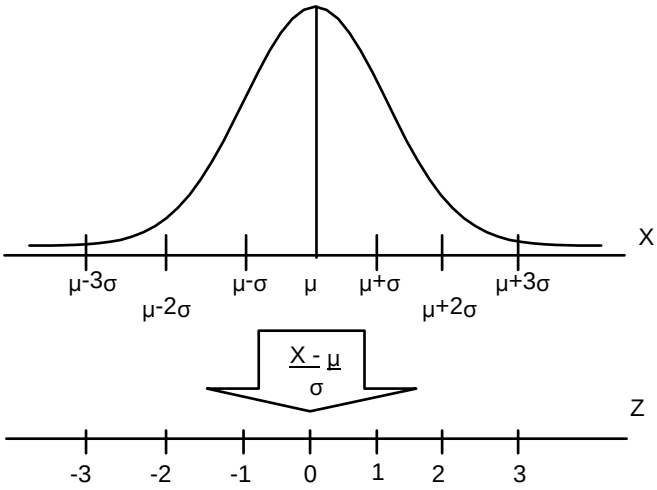


Figura 6 - Transformação de uma N(μ ,σ^2) para uma N(0,1)

A área à esquerda de um valor especificado da N(0,1) encontra-se tabelada.

Utilizando-se a transformação acima, podemos obter as probabilidades para qualquer N(μ ,σ^2). O procedimento é ilustrado através do exemplo abaixo.

Exemplo:

Extrudados tubulares possuem tensão de escoamento (tensão a partir da qual o material se deforma plasticamente), que segue uma distribuição normal com média de 210 MPa com desvio padrão de 5 MPa. Em notação estatística, X ~ N(210 ,5^2). É desejado que tais extrudados tenham tensão de escoamento de pelo menos 200 MPa. Portanto, a probabilidade do extrudado não atingir a especificação desejada é:

Solução:

$$P(X < 200) = P\left(Z < \frac{200 - 210}{5}\right) = P(Z < -2).$$

A figura 7 mostra a transformação realizada e a área desejada.

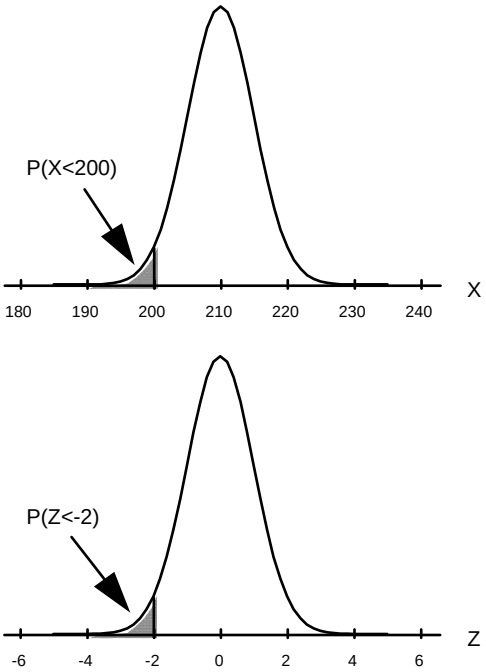


Figura 7 - Probabilidade do extrudado não atingir a especificação desejada

Para cálculo dessa probabilidade, utilizamos a tabela de distribuição normal padronizada (que esta no apêndice do livro indicado na bibliografia básica). Observe que a tabela traz apenas a P(Z<z) para z não negativo (z ≥ 0). As propriedades que se seguem podem ser deduzidas a partir da simetria da densidade em relação à média 0, e são úteis na obtenção de outras áreas não tabuladas.

- P(Z>z) = 1 - P(Z<z)
- P(Z<-z) = P(Z>z)
- P(Z>-z) = P(Z<z)

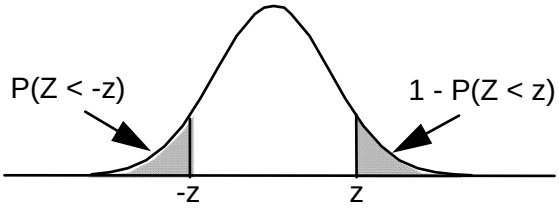


Figura 8 - Áreas correspondentes na distribuição normal

Utilizando as relações apresentadas acima, a probabilidade do extrudado não atender à especificação é

$$P(X < 200) = P(Z < -2) = P(Z > 2) = 1 - P(Z < 2)$$

que, através da tabela da N(0,1) é igual a

$$P(X < 200) = 1 - 0,97725 = 0,02275.$$
$$= 2,275\%$$

Inferência Estatística

Trata-se do processo de obter informações sobre uma população a partir de resultados observados na amostra.

De modo geral, tem-se uma população com grande número de elementos e deseja-se, a partir de uma amostra dessa população, conhecer "o mais próximo possível" algumas características da população.

Toda conclusão tirada por uma amostragem, quando generalizada para a população, virá acompanhada de um grau de incerteza ou risco.

Ao conjunto de técnicas e procedimentos que permitem dar ao pesquisador um grau de confiabilidade, de confiança nas afirmações que faz para a população, baseadas nos resultados das amostras, damos o nome de Inferência Estatística.

O problema fundamental da Inferência Estatística, portanto, é medir o grau de incerteza ou risco dessas generalizações. Os instrumentos da Inferência Estatística permitem a viabilidade das conclusões por meio de afirmações estatísticas.

População e amostra; Estatísticas e parâmetros;

Distribuições amostrais

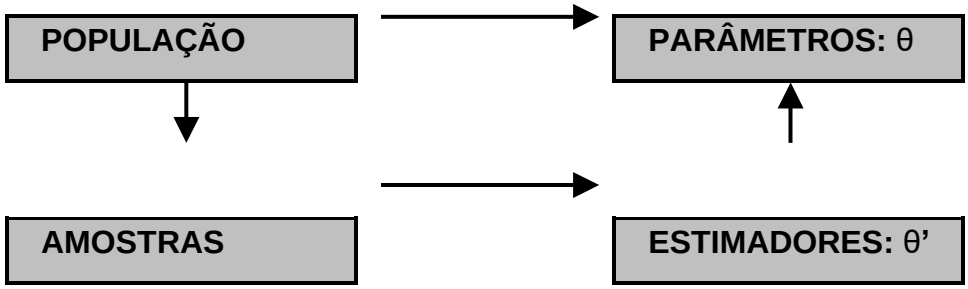
Se um conjunto de dados consiste de todas as observações possíveis (concebíveis ou hipotéticas), é chamado uma população; se um conjunto de dados se consiste apenas de uma parte dessas observações, é chamado uma amostra.

Um dos principais objetivos da maioria dos estudos, análises ou pesquisas estatísticas é fazer generalizações seguras - com base em amostras – em relação às populações das quais se extraíram as amostras.

Definições

Parâmetro: é a medida usada para escrever uma característica numérica populacional. Genericamente é representado por θ. A média (μ), a variância (σ^2) e o coeficiente de correlação (ρ) são alguns exemplos de parâmetros populacionais.

Estimador: também denominado estatística de um parâmetro populacional. É uma característica numérica determinada na amostra, uma função de seus elementos. Genericamente, é representado por θ'. A média amostral (x) e a variância amostral (s^2) são alguns dos exemplos de estimadores.



Distribuição Amostral

Considere todas as possíveis amostras de tamanho n que podem ser extraídas de determinada população. Se para cada uma delas se calcular um valor do estimador, tem-se uma distribuição amostral desse estimador. Como o estimador é uma variável aleatória, pode-se determinar suas características, isto é, encontrar sua média, variância, desvio-padrão.

As distribuições amostrais são fundamentais para o processo de inferência estatística.

Distribuição amostral da Média

Sabe-se que $x' = \frac{\sum x_i}{n}$ (média aritmética) é um estimador da média populacional μ.

O estimador x' é uma variável aleatória; portanto, busca-se conhecer sua distribuição de probabilidade.

Teorema 1 – A média da distribuição amostral das médias, denotada por μ(x'), é igual à média populacional μ.

$$E(x') = \mu(x') = \mu$$

Assim, é provado que a média das médias amostrais é igual à média populacional.

Teorema 2 – Se a população é infinita, ou se a amostragem é com reposição, então a variância da distribuição amostral das médias, denotada por $\sigma^2(x')$, é dada por:

$$VAR\ (x') = \sigma^2_{x'} = \frac{\sigma}{n}^2$$

Teorema 3 – Se a população é finita, ou se a amostragem é sem preposição, então a variância da distribuição amostral das médias é dada por:

$$\sigma^2(x') = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \qquad \sigma_{x'} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Teorema 4 - Se a população tem ou não distribuição normal com média μ e variância σ^2, então a distribuição das médias amostrais será normalmente distribuída com σ^2 média μ e variância $\frac{\sigma^2}{n}$.

Esses quatro teoremas provam que a média amostral (x') tem distribuição normal σ^2 com média igual à média da população (μ) e variância dada por para populações infinitas, assim como $\frac{\sigma^2}{n}$ $(\frac{N-n}{N-1})$ para populações finitas. Ou, ainda:

$$x' \approx N\left(\mu; \frac{\sigma^2}{n}\right) \text{ ou } x' \approx N\left(\mu; \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right)$$

com distribuições padronizadas dadas por:

$$Z_i = \frac{x_i - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \text{ou} \quad Z_i = \frac{\bar{x}_i - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}}$$

Exemplo:

Temos uma população de 5000 alunos de uma faculdade. Sabemos que a altura média dos alunos é de 175 cm e o desvio padrão, de 5 cm. Retiramos uma amostra sem reposição, de tamanho n = 100. Qual o valor do desvio padrão amostral?

Solução:

$$X : N\ (175\ ,\ 25\ cm) \quad \begin{cases} \mu = 175cm \\ \sigma = 5cm \end{cases}$$

Então $\mu_{\bar{x}} = E(\bar{x}) = 175$

E $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{5}{10}\sqrt{\frac{5000-25}{5000-1}} = 0{,}4988$

Estimação

Há dois tipos fundamentais de estimação: por ponto e por intervalo.

Estimação pontual

O problema da estimação pontual surge quando estamos interessados em alguma característica numérica de uma distribuição desconhecida (ex: média, variância) e desejamos calcular, a partir de observações, um número que inferimos que seja uma aproximação da característica numérica em questão.

Para ilustrar alguns dos problemas com os quais nos deparamos quando estimamos a média de uma população com base em dados amostrais, vamos recorrer a um estudo em que planejadores industriais procuraram determinar o tempo médio que um adulto leva para montar um robô “fácil de montar”. Com uma amostra aleatória, obtém-se os seguintes dados (em minutos) para 36 pessoas que montaram o robô:

17	13	18	19	17	21	29	22	16	28
21	15	26	23	24	20	8	17	17	21
32	18	25	22	16	10	20	22	19	14
30	22	12	24	28	11				

A média desta amostra é x' = 19,9 minutos. Na ausência de qualquer outra informação, podemos tomar esta cifra como uma estimativa de μ, o "verdadeiro" tempo médio que um adulto leva para montar o robô.

Esse tipo de estimativa é chamada estimativa pontual, pois consiste de um único número, ou um único ponto na escala dos números reais. Embora se trate da forma mais comum de expressar estimativas, ela deixa margem para não poucas questões. Por exemplo, não nos diz em quantas informações a estimativa se baseia, nem tampouco nos informa sobre o tamanho possível do erro.

Estimação por intervalo

A estimação por pontos de um parâmetro não possui uma medida do possível erro cometido na estimação, daí surge a ideia de construir os intervalos de confiança, que são baseados na distribuição amostral do estimador pontual.

Uma maneira de expressar a precisão da estimação é estabelecer limites que, com certa probabilidade, incluam o verdadeiro valor do parâmetro da população. Esses limites são chamados "limites de confiança": determinam um intervalo de confiança, no qual deverá estar o verdadeiro valor do parâmetro. Logo, a estimação por intervalo consiste na fixação de dois valores tais que (1 - α) seja a probabilidade de que o intervalo, por eles determinado, contenha o verdadeiro valor do parâmetro.

α : nível de incerteza ou grau de desconfiança

1 - α : coeficiente de confiança ou nível de confiabilidade.

Portanto, α nos dá a medida da incerteza desta inferência (nível de significância).

Logo, a partir das informações de amostra, devemos calcular os limites de um intervalo, valores críticos que em (1 - α)% dos casos inclua o valor do parâmetro a estimar e em α% dos casos não inclua o valor do parâmetro.

Intervalo de confiança (IC) para a média populacional (μ) quando a Variância (σ^2) é conhecida.

Como se sabe, o estimador de μ é x'. Também é conhecida a distribuição de probabilidade de x':

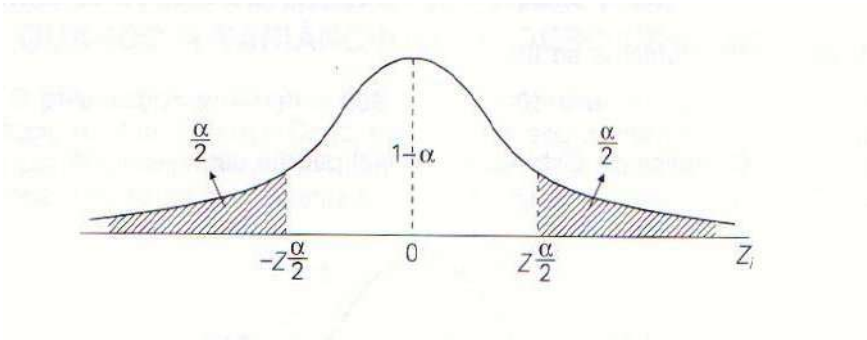
$$x' \approx N\left(\mu; \frac{\sigma^2}{n}\right)$$ para as populações infinitas,

$$x' \approx N\left(\mu; \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right)$$ para as populações finitas.

Assim, para o caso de populações infinitas, a variável padronizada de x' será:

$$Z = \frac{x' - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Fixando-se um nível de confiança 1 - α, tem-se:



Ou seja:

$$P\left(-Z\frac{\alpha}{2} \le Z \le Z\frac{\alpha}{2}\right) = 1 - \alpha$$

Substituindo-se o valor de Z, tem-se:

$$P\left(-Z\frac{\alpha}{2} \le \frac{x'-\mu}{\frac{\sigma}{\sqrt{n}}} \le Z\frac{\alpha}{2}\right) = 1-\alpha$$

Resolvendo-se as duas inequações para μ, tem-se o intervalo de confiança para a média populacional (μ) quando a variância (σ^2) é conhecida:

$$P\left(x'-Z\frac{\alpha}{2}\frac{\sigma}{\sqrt{n}} \le \mu \le x'+Z\frac{\alpha}{2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Como poderá ser verificado, a aplicação da fórmula é extremamente simples. Fixase o valor de 1 - α , ou (1 - α)100= %, e observa-se na tabela de distribuição normal padrão o valor das abscissas que deixam α/2 em cada uma das caudas. Com os valores de x' (média amostral), σ=desvio padrão da população, que neste caso é conhecido, e n (tamanho da amostra), constrói-se o intervalo.

Para o caso de populações finitas, usa-se a seguinte fórmula:

$$P\left(x'-Z\frac{\alpha}{2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \le \mu \le x'+Z\frac{\alpha}{2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right) = 1-\alpha$$

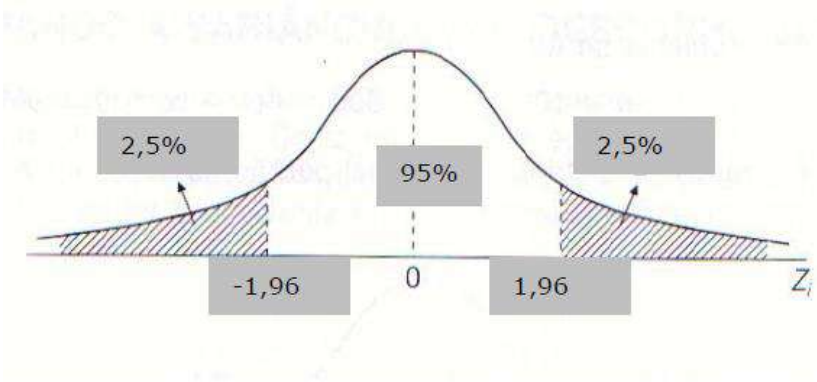
Exemplo:

A duração da vida de uma peça de equipamento é tal que σ=5 horas. Foram amostradas 100 dessas peças, obtendo-se a média de 500 horas. Deseja-se construir um intervalo de confiança para a verdadeira duração média da peça com um nível de 95%.

Solução:

σ = 5 ; n = 100 x'=500 (1 - α)100=95%

o gráfico da distribuição normal padrão será:



lembre-se que para descobrir a abscissa 1,96, entrou-se na tabela de distribuição normal padronizada com o valor 0,475 = 47,5 , já que a tabela é de faixa central.

Substituindo na formula :

$$P\left(x'-Z\frac{\alpha}{2}\frac{\sigma}{\sqrt{n}} \le \mu \le x'+Z\frac{\alpha}{2}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

$$P\left(500-1{,}96\frac{5}{\sqrt{100}} \le \mu \le 500+1{,}96\frac{5}{\sqrt{100}}\right)=95\%$$

Efetuando os cálculos temos:

P(499,02 ≤ μ ≤ 500,98) = 95%

Testes de Hipóteses

Trata-se de uma técnica para se fazer inferência estatística. Ou seja, a partir de um teste de hipóteses realizado com os dados amostrais, pode-se fazer inferências sobre a população.

Principais conceitos

Hipóteses Estatística – Trata-se de uma suposição quanto ao valor de um parâmetro populacional, ou quanto à natureza da distribuição de probabilidade de uma variável populacional. São exemplos de hipóteses estatísticas:

a) Os chips da marca A têm vida média H :$\mu = \mu_0$;

b) O nível de inteligência de uma população de universitários é H : $\mu = \mu_0$;

c) O aço produzido pelo processo A é mais duro que o aço produzido pelo processo B:$\mu_A > \mu_B$;

d) A altura média da população brasileira é de 1,65m, isto é: H :μ=1,65m;

e) A variância populacional dos salários vale R$ 5.000^2, isto é: H :$\sigma^2=5.000^2$;

f) A proporção de paulistas com a doença X é de 40 %, ou seja: H : p=0,40;

g) A distribuição dos pesos dos alunos da nossa faculdade é normal;

h) A chegada de navios ao porto de Santos é descrita por uma distribuição de Poisson.

Formulamos duas hipóteses básicas:

H_0: hipótese nula ou da existência;

H_1: hipótese alternativa.

Testamos hipóteses para tomarmos uma decisão entre duas alternativas. Por essa razão, o Teste de Hipótese é um Processo de Decisão Estatística.

Tipos de Hipótese – Designa-se por H_0, chamada hipótese nula, a hipótese estatística a ser testada, e por H_1 a hipótese alternativa. A hipótese nula expressa uma igualdade, enquanto a hipótese alternativa é dada por uma desigualdade.

Exemplo:

$$\left.\begin{array}{l} H_0 : \mu = 1{,}65m \\ H_1 : \mu \neq 1{,}65m \end{array}\right\}$$

- para testes bilaterais (dá origem a um teste bicaudal)

$$\left.\begin{array}{l} H_0 : \mu = 1{,}65m \\ H_1 : \mu > 1{,}65m \end{array}\right\}$$

- para testes unilaterais à direita (dá origem a um teste unilateral à direita) H_1 :

$$\left.\begin{array}{l} H_0 : \mu = 1{,}65m \\ H_1 : \mu < 1{,}65m \end{array}\right\}$$

- para testes unilaterais à esquerda (dá origem a um teste unicaudal à esquerda)

O procedimento padrão para a realização de um Teste de Hipóteses é o seguinte:

- Define-se as hipóteses do teste: nula e alternativa;
- Fixa-se um nível de significância α;
- Levanta-se uma amostra de tamanho n e calcula-se uma estimativa θ'_0 do parâmetro θ;
- Usa-se para cada tipo de teste uma variável cuja distribuição amostral do estimador dos parâmetros seja a mais concentrada em torno do verdadeiro valor do parâmetro;
- Calcula-se com o valor do parâmetro θ_0, dado por H_0, o valor crítico, valor observado na amostra ou valor calculado (V_{calc});
- Fixa-se duas regiões: uma de não rejeição de H_0 (RNR) e uma de rejeição de H_0 ou crítica (RC) para o valor calculado, ao nível de risco dado;
- Se o valor observado (V_{calc}) ∈ Região de Não Rejeição, a decisão é a de não rejeitar H_0;
- Se V_{calc} ∈ Região Crítica, a decisão é a de rejeitar H_0.

Devemos observar que quando se fixa α , determinamos para os testes bilaterais, por exemplo, valores críticos (tabelados), Vα, tais que:

$$\mathbf{P(|V_{calc}| < V\alpha) = 1 - \alpha \rightarrow RNR}$$

$$\mathbf{P(|V_{calc}| \supseteq V_{\alpha}) = \alpha \rightarrow RC}$$

Testes de Hipóteses para a Média de Populações normais com variâncias (σ^2) conhecidas

Testes Bilaterais

De uma população normal com variância 36, toma-se uma amostra casual de tamanho 16, obtendo-se x'=43. Ao nível de 10%, testar as hipóteses:

$$\begin{cases} H0 : \mu = 45 \\ H1 : \mu \neq 45 \end{cases}$$

Solução: Como o teste é para média de populações normais com variância conhecida, usaremos a variável Z: N(0,1) como critério.

σ^2=36 **x'=43** **n=16**

$$Z = \frac{x' - \mu_{H0}}{\sigma_{x'}} \quad , \quad \sigma_{x'} = \frac{\sigma}{\sqrt{n}}$$

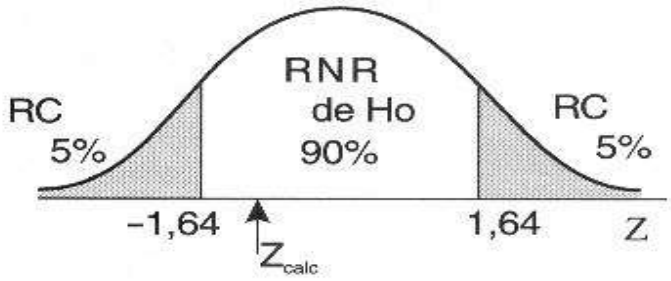
Como o teste é bilateral e α =10%, a Região de Não Rejeição, (RNR), é:

P(|Z|< Z_α)=1 - α → P(|Z|<1,64)=0,90

Zα=Z5%=1,64

E a Região de Rejeição (RC) é dada por

P(|Z| ≥Z_α)=α→P(|Z|≥1,64) =0,10



Como Z_{calc}=-1,33

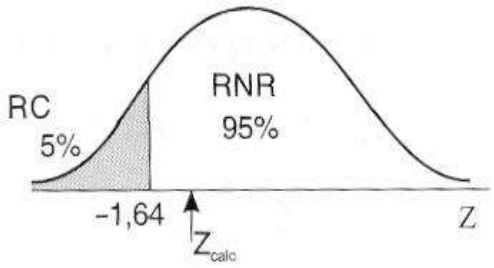
Temos que $Z_{calc} \in$ RNR

Logo, a decisão é não rejeitarmos H_0, isto é, a média é de 45, com 10% de risco de não rejeitarmos uma hipótese falsa.

Testes Unilateral (Monocaudal) à Esquerda

Uma fábrica anuncia que o índice de nicotina dos cigarros da marca X apresenta-se abaixo de 26 mg por cigarro. Um laboratório realiza 10 análises do índice obtendo: 26,24,23,22,28,25,27,26,28,24. Sabe-se que o índice de nicotina dos cigarros da marca X se distribui normalmente com variância 5,36 mg2. Pode-se aceitar a afirmação do fabricante, ao nível de 5%?

$$\begin{cases} H0 : \mu = 26 \\ H1 : \mu < 26 \end{cases}$$



∴RNR= (-1,64; +∞)

∴RC = (-∞ ; -1,64]

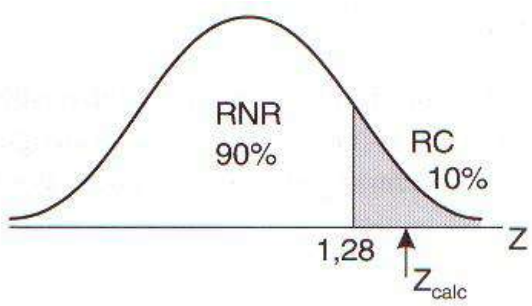
∴Z_{calc} ∈ RNR

Não se rejeita H_0, isto é, ao nível de 5% podemos concluir que a afirmação do fabricante é falsa.

Testes Unilateral (Monocaudal) à Direita

Um fabricante de lajotas de cerâmica introduz um novo material em sua fabricação e acredita que aumentará a resistência média, que é de 206 kg. A resistência das lajotas tem distribuição normal, com desvio padrão de 12 kg. Retira-se uma amostra de 30 lajotas, obtendo-se X' =210 kg. Ao nível de 10%, pode o fabricante aceitar que a resistência média de suas lajotas tenha aumentado?

$$\begin{cases} H0: \mu = 206 \\ H1: \mu > 206 \end{cases}$$



∴RNR=(-∞ ; 1,28)

∴RC = [1,28; +∞)

∴Z_{calc} ∈ RC

Como $Z_{calc} > Z\alpha$, rejeita-se H_0, isto é, ao nível de 10% o fabricante pode concluir que a resistência média de suas lajotas aumentou.

Erros de Decisão

Tipos de Erro – Há dois tipos possíveis de erro ao testar uma hipótese estatística. Pode-se rejeitar uma hipótese quando ela é, de fato, verdadeira, ou aceitar uma hipótese quando ela é, de fato, falsa. A rejeição de uma hipótese verdadeira é chamada “erro tipo I”. A aceitação de uma hipótese falsa constitui um “erro tipo II”.

As probabilidades desses dois tipos de erros são designadas, respectivamente, por α e β.

A probabilidade α do erro tipo I é denominada “nível de significância” do teste.

Resumindo:

		Realidade	
		H_0 verdadeira	H_0 falsa
Decisão	Aceitar H_0	Decisão Correta (1-α)	Erro tipo II (β)

	Rejeitar H_0	Erro tipo I (α)	Decisão Correta (1 - β)

Observe que o erro tipo I só poderá ser cometido se se rejeitar H_0; e o erro do tipo II, quando se aceitar H_0.

O objetivo, obviamente, é reduzir ao mínimo as probabilidades dos dois tipos de erros. Infelizmente, essa é uma tarefa difícil porque, para uma amostra de determinado tamanho, a probabilidade de se incorrer em um erro tipo II aumenta à medida que diminui a probabilidade do erro I. E vice-versa. A redução simultânea dos erros poderá ser alcançada pelo aumento do tamanho da amostra.

Correlação e Regressão linear

Em muitas situações, torna-se interessante e útil estabelecer uma relação entre duas ou mais variáveis. A matemática estabelece vários tipos de relações entre variáveis, por exemplo, as relações funcionais e as correlações.

Relações Funcionais

São relações matemáticas expressas por sentenças matemáticas, cujos exemplos apresentamos a seguir:

- Área do retângulo (A=a.b) é a relação entre os lados do retângulo;
- Densidade de massa (d_m= m/v) é a relação entre a massa e o volume de um corpo;
- Perímetro de uma circunferência (C=2πR) é a relação entre o comprimento da circunferência e o valor do raio.

Relações Estatísticas e Correlações

São relações estabelecidas após uma pesquisa. Com base nos resultados da pesquisa, são feitas comparações que eventualmente podem conduzir (ou não) à ligação entre as variáveis.

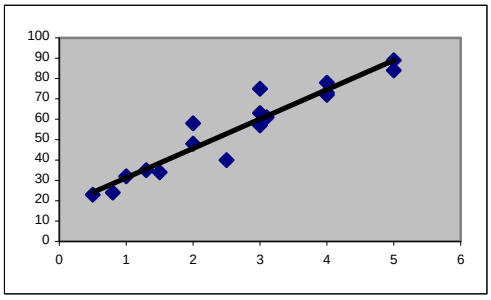
Exemplo: relação entre a idade e a estatura de uma criança, ou a relação entre a classe social de uma pessoa e o número de viagens por ela realizado.

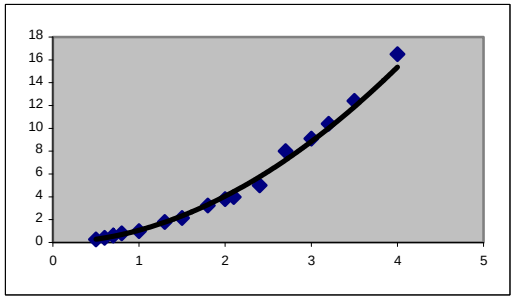
No estudo estatístico, a relação entre duas ou mais variáveis denomina-se correlação. A utilidade e importância das correlações entre duas variáveis podem conduzir à descoberta de novos métodos, cujas estimativas são vitais em tomadas de decisões.

Diagrama de dispersão

O diagrama de dispersão é um gráfico cartesiano em que cada um dos eixos corresponde às variáveis correlacionadas. A variável dependente (Y) situa-se no eixo vertical e o eixo das abscissas é reservado para a variável independente (X). Os pares ordenados formam uma nuvem de pontos.

A configuração geométrica do diagrama de dispersão pode estar associada a uma linha reta (correlação linear), uma linha curva (correlação curvilínea) ou, ainda, ter os pontos dispersos de maneira que não definam nenhuma configuração linear; nesta última situação, não há correlação.



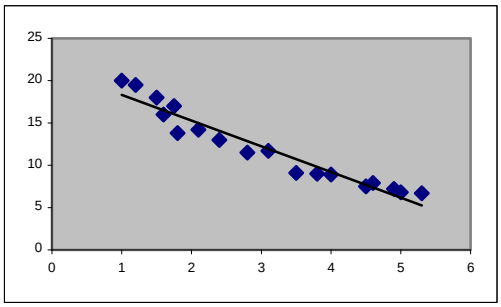


Correlação Linear Correlação Curvilínea

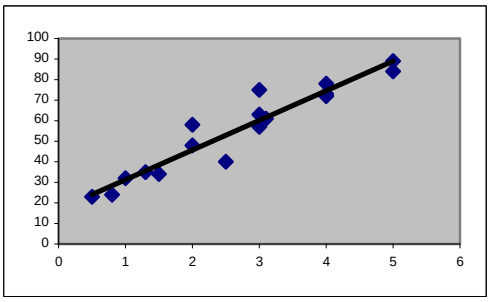
Figura 9.1. Diagramas de dispersão

Correlação Linear

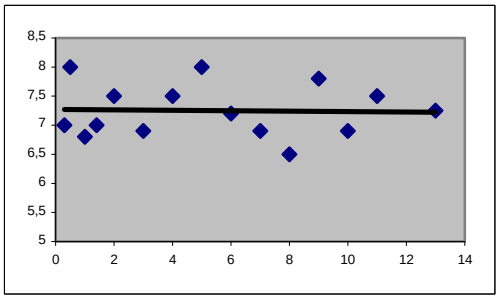
Correlação linear é uma correlação entre duas variáveis, cujo gráfico aproxima-se de uma linha. É uma linha de tendência, porque procura acompanhar a tendência da distribuição de pontos, que pode corresponder a uma reta ou a uma curva. Por outro lado, é, também, uma linha média, porque procura deixar a mesma quantidade de pontos abaixo e acima da linha.



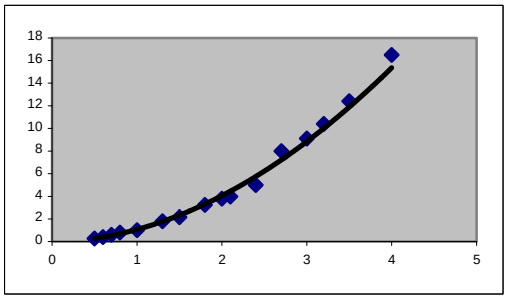
Correlação Linear positiva



Correlação Linear negativa



Não há correlação



Relação curvilínea direta

Figura 9.2. Diagramas de dispersão de diversos tipos de correlação.

Para definir se a correlação entre as variáveis corresponde a uma linha reta ou a uma curva, pode-se utilizar modos qualitativos ou quantitativos.

No modo qualitativo, vai imperar o “bom senso” do pesquisador para verificar qual o grau de intensidade na correlação entre as variáveis; isso significa o estabelecimento de uma relação numérica que medirá o nível da correlação.

Coeficiente de Correlação Linear (r)

O coeficiente de correlação linear pode ser apresentado como uma medida de correlação, pois tem como objetivo indicar o nível de intensidade que ocorre na correlação entre as variáveis. O coeficiente de correlação linear pode ser positivo ou negativo. O sinal positivo do coeficiente de correlação linear indica que o sentido da correlação corresponde a uma reta de inclinação descendente, e o sinal negativo corresponde a uma reta de inclinação ascendente. Uma das formas de medir o coeficiente de correlação linear foi desenvolvido por Pearson e recebe o nome de coeficiente de correlação de Pearson. O coeficiente de correlação de Pearson mede o grau de ajustamento dos valores em torno de uma reta.

Coeficiente de Correlação de Pearson (r):

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(y_i\right)}{\sqrt{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right] * \left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]}}$$

Temos

r = o coeficiente de Pearson

n = o número de observações

x_i = variável independente

y_i =variável dependente

O valor do coeficiente de correlação r tem a variação entre +1 e –1, ou seja, está limitado entre os valores do Intervalo[-1,+1].

- r = +1 (correlação positiva entre as variáveis);
- r = - 1 (correlação perfeita negativa entre as variáveis);
- r = 0 (não há correlação entre as variáveis ou, ainda, a correlação não é linear, caso exista).

Quanto mais próximo o valor de r estiver do valor “1”, mais forte a correlação linear.

Quanto mais próximo o valor de r estiver do valor “0”, mais fraca a correlação linear.

Em geral, multiplica-se o valor de r por 100; dessa forma, o resultado passa a ser expresso em porcentagem. Na prática, estabelecem-se critérios para verificar os diversos níveis do fraco ao forte, chegando até o perfeito:

- 0<|r|<0,3: a correlação é fraca e fica difícil estabelecer relação entre as variáveis. Em porcentagem: 0<|r|< 30%;
- 0,3≤|r|< 0,6: a correlação é fraca, porém, podemos considerar a existência de relativa correlação entre as variáveis. Em porcentagem: 30%≤|r| <60%;
- 0,6≤|r| <1: a correlação é de média para forte; a relação entre as variáveis é significativa, o que permite coerência com poucos conflitos na obtenção das conclusões. Em porcentagem: 60%≤|r|≤100%.
-

Exemplo:

Uma pesquisa pretende verificar se há correlação significativa entre o peso total do lixo descartado, por dia, numa empresa com o peso do papel contido nesse lixo.

Hotel	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}
Peso total	10,47	19,85	21,25	24,36	27,38	58,09	33,61	35,75	38,33	49,14
Peso do papel	2,43	5,12	6,88	6,22	8,84	8,76	7,54	8,47	9,55	11,43

De acordo com os dados, fazemos a representação gráfica. Os pares ordenados formam o diagrama de dispersão.

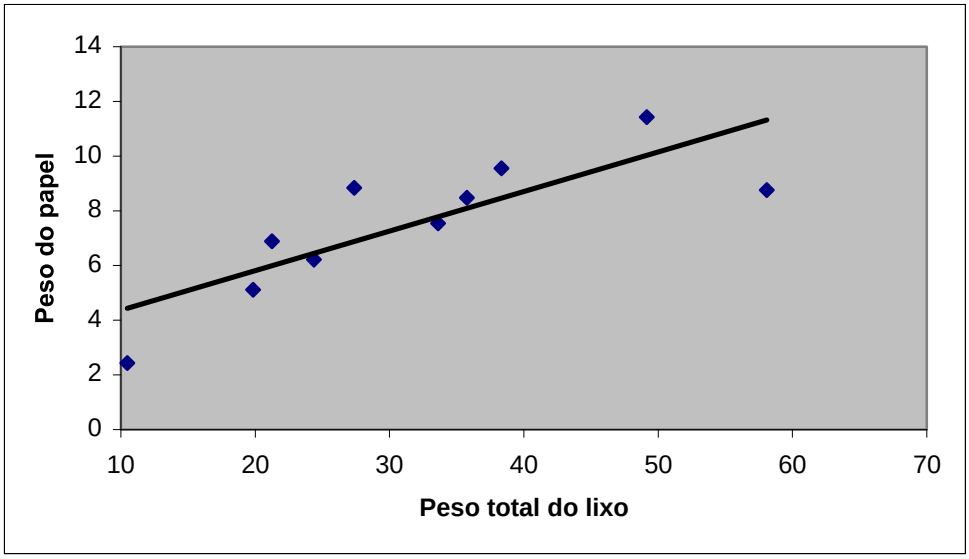


Figura 9.3. Correlação entre o peso total do lixo descartado e o peso do papel contido nesse lixo

Para se verificar o grau de correlação entre as variáveis, calcula-se o coeficiente de correlação linear pela fórmula do coeficiente de correlação de Pearson:

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(y_i\right)}{\sqrt{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right] * \left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]}}$$

	Peso total (x_i)	Peso papel (y_i)	$x_i\ y_i$	$x_i 2$	$y_i 2$
H1	10,47	2,43	25,44	109,62	5,90
H2	19,85	5,12	101,63	394,02	26,21
H3	21,25	6,88	146,20	451,56	47,33
H4	24,36	6,22	151,52	593,41	38,69
H5	27,38	8,84	242,04	749,66	78,15
H6	28,09	8,76	246,07	789,05	76,74
H7	33,61	7,54	253,42	1129,63	56,85

H8	35,73	8,47	302,63	1276,63	71,74
H9	38,33	9,55	366,05	1469,19	91,20
H10	49,14	11,43	561,67	2414,74	130,64
Σ	288,21	75,24	2396,68	9377,52	623,47

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right] * \left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]}}$$

$$r = \frac{10 * 2396{,}68 - 288{,}21 * 75{,}24}{\sqrt{[10 * 9377{,}52 - (288{,}21)^2] * [10 * 623{,}47 - (75{,}24)^2]}}$$

$$r = \frac{23966{,}8 - 21684{,}9}{\sqrt{[93775{,}2 - 83065] * [6234{,}7 - 5661{,}1]}} = \frac{2281{,}83}{\sqrt{10 * 710{,}21 * 573{,}59}} = \frac{2281{,}83}{2478{,}57} = 0{,}9206$$

Observamos, assim: 0,6 ≤ r ≤1. Esse resultado indica que há uma forte correlação entre as variáveis ou, ainda, que a correlação entre as duas variáveis é bastante significativa. Nesse caso, podemos concluir haver coerência na afirmação de que existe correlação entre o peso total do lixo descartado e o peso do papel contido nesse lixo.

Regressão – Reta de regressão (ou reta de mínimos quadrados ou reta de ajuste)

A correlação linear é uma correlação entre duas variáveis, cujo gráfico aproxima-se de uma linha. O gráfico cartesiano que representa essa linha é denominado diagrama de dispersão. Para poder avaliar melhor a correlação entre as variáveis, é interessante obter a equação da reta; essa reta é chamada de reta de regressão e a equação que a representa é a equação de regressão. O diagrama de

dispersão é construído de acordo com os dados amostrais de n observações e a equação de regressão é dada pela expressão:

$$\mathbf{Y= aX + b \ \rightarrow Y' = aX + b}$$

X é a variável independente

Y→Y" é a variável dependente; na verdade, é a variável correlacionada com a variável X e sobre a qual se obtém um valor estimado.

Esse tipo de notação, de Y para Y', caracteriza que não se trata de uma relação funcional para a determinação da reta, e sim de uma relação estatística, em que a distribuição está baseada em estimativas de dados colhidos por amostragem.

Sendo a e b os parâmetros de equação da reta, esses podem ser calculados por meio das fórmulas:

$$a = \frac{n\sum x_i y_i - \sum x_i * \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad b = \bar{y} - a\bar{x}$$

Sendo:

n = número de observações dos dados amostrais

y = valor médio da variável y; o cálculo faz-se pela expressão $\bar{y} = \frac{\sum y_i}{n}$

x = valor médio da variável x; o cálculo faz-se pela expressão $\bar{x} = \frac{\sum x_i}{n}$

Exemplo:

Determine a equação da reta de regressão do exemplo anterior, que trata de uma pesquisa entre o peso total do lixo descartado por dia com o peso do papel contido nesse lixo.

Para a obtenção da equação da reta de regressão, elabora-se inicialmente uma tabela contendo nas colunas as variáveis dependentes (y_i), as independentes (x_i) e os produtos $x_i y_i$ e x_i^2.

Cálculo do parâmetro a da equação da reta:

$$a = \frac{n\sum x_i y_i - \sum x_i * \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{10 * 2396{,}68 - 288 * 75{,}24}{10 * 9377{,}52 - (288{,}21)^2} = \frac{23996{,}8 - 21684{,}9}{93775{,}2 - 83065}$$

$$a = \frac{2281{,}83}{10710{,}2} = 0{,}213$$

Cálculo do parâmetro b da equação da reta:

$$\bar{y} = \frac{75{,}24}{10} = 7{,}52 \qquad \text{e} \qquad \bar{x} = \frac{288{,}21}{10} = 28{,}82$$

$$b = \bar{y} - a\bar{x} = 7{,}52 - 0{,}213 * 28{,}82 = 7{,}52 - 6{,}14 = 1{,}38$$

Uma vez calculados os parâmetros a e b, pode-se escrever a equação da reta:

Y' = 0,213 X + 1,38

Para o traçado de uma reta, basta que se conheça dois de seus pontos. Assim, com base na equação da reta acima, pode-se estabelecer dois pontos para X e Y'.

- Para X = 0, temos Y' = 1,38
- Para X = 50, temos Y' = 12,03
-

De acordo com os pontos $P_1(0;1{,}38)$ e $P_2(50;12{,}03)$, pode-se traçar a reta de regressão.

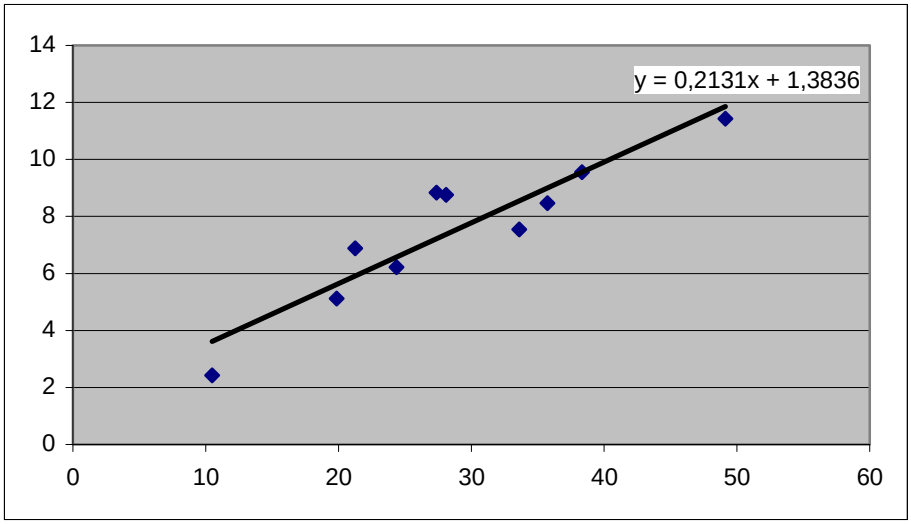


Figura 9.4. Correlação entre o peso total do lixo descartado e o peso do papel contido nesse lixo

Com base no conhecimento da equação da reta, pode-se interpolar e extrapolar valores.

- Interpolação: a interpolação ocorre quando o valor considerado pertence ao intervalo da tabela, porém, não figura entre os dados coletados.

Supondo-se o valor 15 kg para o peso total do lixo descartado, pode-se estimar o peso de papel contido nesse lixo. Uma vez que 15 kg não é um dado coletado e, consequentemente, não pertence à tabela de dados, utiliza-se a equação da reta para determinar o valor correspondente ao peso do papel.

Para 15 kg de lixo descartado, estima-se que haja 4,58 kg de papel contido nesse lixo.

- Extrapolação: a extrapolação ocorre quando o valor considerado não pertence ao intervalo da tabela, e também não figura entre os dados coletados.

Suponha que o peso do lixo descartado seja de 60 kg. Esse valor não é um dado coletado e nem se encontra dentro do intervalo [10,47, 49,14]. Essa situação é semelhante à anterior e utiliza-se a equação de reta para determinar o peso do papel.

Para 60 kg de lixo descartado, estima-se, por extrapolação, que haja 14,16 kg de papel contido nesse lixo.

Referências

BUSSAB, W.O. e Morettin, P.A. Estatística Básica. São Paulo: Atual, 1987.

FONSECA, J.S. e Martins, G.A. Curso de Estatística. São Paulo: Atlas, 1993.

LAPPONI, J.C. Estatística usando Excel 5 e 6. São Paulo: Lapponi Treinamento e Editora, 1997.

MORETTIN, L.G. Estatística Básica – Vol. 2 – Inferência. São Paulo: Makron Books, 1999.

MORETTIN, L.G. Estatística Básica – Vol.1 – Probabilidade. São Paulo: Makron Books, 1999.

STEVENSON,W.J. Estatística Aplicada à Administração. São Paulo: Harbra, 1996.

TIBONI,C.G. R. Estatística Básica para o curso de Turismo. São Paulo: Atlas, 2002.

TOLEDO, G. L. e Ovalle, I.I. Estatística Básica. São Paulo: Atlas, 1985.