Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências Agrárias e Veterinárias
Departamento de Ciências Agrárias

Estatística
e
Bioestatística

Disciplinas ministradas aos cursos de graduação em Agronomia, Ciência Biológicas e Zootecnia.

2019

Sumário

1 ESTATÍSTICA E BIOESTATÍSTICA

1.1 Introdução a Estatística

1.1.1 Porque estudar Estatística?

O nome, estatística, é derivado da palavra latina "status". Originalmente essa palavra significava "informações úteis ao Estado" (para fins de taxação, conhecimentos dos recursos do país, da composição da população entre outros). Posteriormente, a palavra passou a significar dados quantitativos que apresentavam tendência de flutuarem de uma forma mais ou menos imprevisível, significado esse que permanece até hoje quando se falam em estatísticas de, por exemplo, acidentes de trabalho, do número de nascimentos ou mortes, etc.

Mais recentemente, a palavra passou a significar a ciência que diz respeito à coleta, organização e análise dos dados quantitativos de tal forma que seja possível efetuar julgamentos racionais sobre os mesmos. A estatística tem também a função de auxiliar do método científico, especialmente no planejamento experimental, na coleta de dados, na interpretação analítica dos experimentos (análise dos dados experimentais) e na estimação dos parâmetros da população. Em alguma fase de um trabalho nos deparamos com o problema de analisar e entender um conjunto de dados relevante ao nosso particular objetivo de estudo. É necessário trabalhar os dados para transformá-los em informações, para compará-los com outros resultados, ou ainda para julgar a adequação de alguma teoria ou hipótese. De modo bem geral, podemos dizer que a essência da **Ciência** é a ***observação*** e que o seu objetivo básico é a ***inferência***.

Além disso, o uso de técnicas computacionais pode parecer um problema para o pesquisador ou estudante cujo treino e interesse não envolva a matemática, entretanto, a estatística é uma realidade na literatura científica e especializada. Então, julgamos razoável que o profissional das áreas de biológicas e agrária adquira um mínimo de conhecimento técnico sobre estatística. Outro resultado do estudo da estatística é a familiarização com os termos técnicos da área, uma vez que a falta de conhecimento de certos termos pode resultar na total incompreensão de um artigo científico, ou de uma exposição de ideias e hipótese de pesquisadores e profissionais que possuem tal conhecimento.

1.1.2 Estatística e Bioestatística

Os pesquisadores de disciplinas relacionadas às ciências biológicas, agrárias e à saúde utilizam uma grande variedade de ferramentas para entende os fenômenos estudados por eles. Uma das mais importantes é a **bioestatística/estatística**, pois esta desempenha um papel fundamental na análise de dados coletados no contesto de testes químicos e ensaios biológicos, bem como em estudos de outras áreas como epidemiologia, política sanitária, saúde pública e familiar entre outras. A **Bioestatística** é um ramo mais amplo da área **Estatística**. Então, para fins didáticos vamos, inicialmente, definir o termo **Estatística**.

A **Estatística** é fundamental na análise de dados provenientes de quaisquer processos onde exista variabilidade, estando assim, interessada nos métodos e processos quantitativos que servem para a coleta, organização, resumo, apresentação e análise desses dados, bem como na obtenção de conclusões válidas e na tomada de decisões a partir de tais análises. Assim, de maneira geral, a estatística pode ser dividida em três áreas:

A **Estatística Descritiva:** geralmente utilizada nas etapas inicias dos trabalhos, se refere à maneira de representar dados em tabelas e gráficos, resumi-los por meio de algumas medidas sem, contudo, tirar quaisquer informações sobre um grupo maior. Portanto, informações e conclusões a respeito do fenômeno estudado são tiradas de modo informal e direto, restritas àquele ***particular*** conjunto de valores.

A **Probabilidade**: é a teoria matemática utilizada para se estudar a ***incerteza*** oriunda de fenômenos de caráter ***aleatório***. Seu estudo é fundamental na **bioestatística/estatística**, tem sua origem ligada aos jogos de azar. Esses jogos implicam em ações como girar uma roleta, lançar um dado ou uma moeda, tendo como característica a ***incerteza*** de ocorrer determinado acontecimento (como a face cara de uma moeda, ou o às de ouro em um set de baralho) em determinada tentativa, e a regularidade em longo prazo, que permite prever o número de vezes que ocorrerá determinado acontecimento em uma série de tentativas conduzidas de maneira uniforme.

A **Inferência Estatística**: ao contrário da estatística descritiva, é o estudo de técnicas que possibilitem a extrapolação das informações e conclusões obtidas a partir de subconjuntos de dados, a um grande número de dados, ou seja, procura estabelecer conclusões para toda uma **população**, quando apenas se observou uma parte desta (**denominada mostra**).

De maneira geral a **Bioestatística** é a **Estatística** aplicada a dados biológicos e de ciências agrárias, como tal, está interessada na coleta, organização, resumo, apresentação e análise de tais dados.

1.2 Conceitos básicos

1.2.1 Populações e amostras

Na terminologia estatística, o grande conjunto de dados que contém a característica que temos interesse recebe o nome de **População**. Esse termo refere-se não semente a uma coleção de indivíduos, mas também ao alvo sobre o qual reside o nosso interesse. Assim, nossa população pode ser tanto todo o conjunto de cervos em uma área de proteção, todas as árvores de uma determinada espécie na floresta amazônica, todas as lâmpadas produzidas em uma fábrica em um determinado período de tempo. Dentro dessa definição de população, poderemos, ainda, fazer uma distinção entre os tipos de população:

Populações Comuns: "*Uma população é um conjunto de pessoas (ou coisas) que possuem uma característica observável comum*" – este é o conceito mais amplo de população, e temos como exemplos: população de pessoas que moram na Região Sudeste do Brasil que apresentam resultado positivo para hepatite C, a população de plantas de uma variedade de soja plantada na região sul do Brasil, a população de bovinos de corte do estado do Mato Grosso do Sul.

Populações Estatísticas: "*a população estatística se refere a dados (informação), e não às pessoas, indivíduos ou objetos*" nessa abordagem, a população é composta de características das pessoas (ou objetos de estudo). Tomando o exemplo anterior, na população comum de pessoas que moram na Região Sudeste do Brasil que apresentam positivo para hepatite C, teríamos como populações estatísticas um parâmetro que indicasse se todas as pessoas necessitaram de transfusão sanguínea em algum momento de suas vidas, por exemplo. No caso da população de uma variedade específica de soja teríamos como população estatística, a sua produtividade. Portanto, a população estatística consiste em ***características*** de pessoas ou objetos de estudo, independente de terem sido medidas ou não.

Amostra: Na maioria dos casos, não conseguimos acessar toda uma população para estudar as características de interesse, isso devido às razões econômicas, éticas e dificuldades de outra natureza. Assim, tomaremos alguns elementos dessa população para formar um grupo a ser estudado. Este subconjunto da população, em geral com menores dimensões, é denominado **amostra**, ou seja, qualquer subconjunto da população.

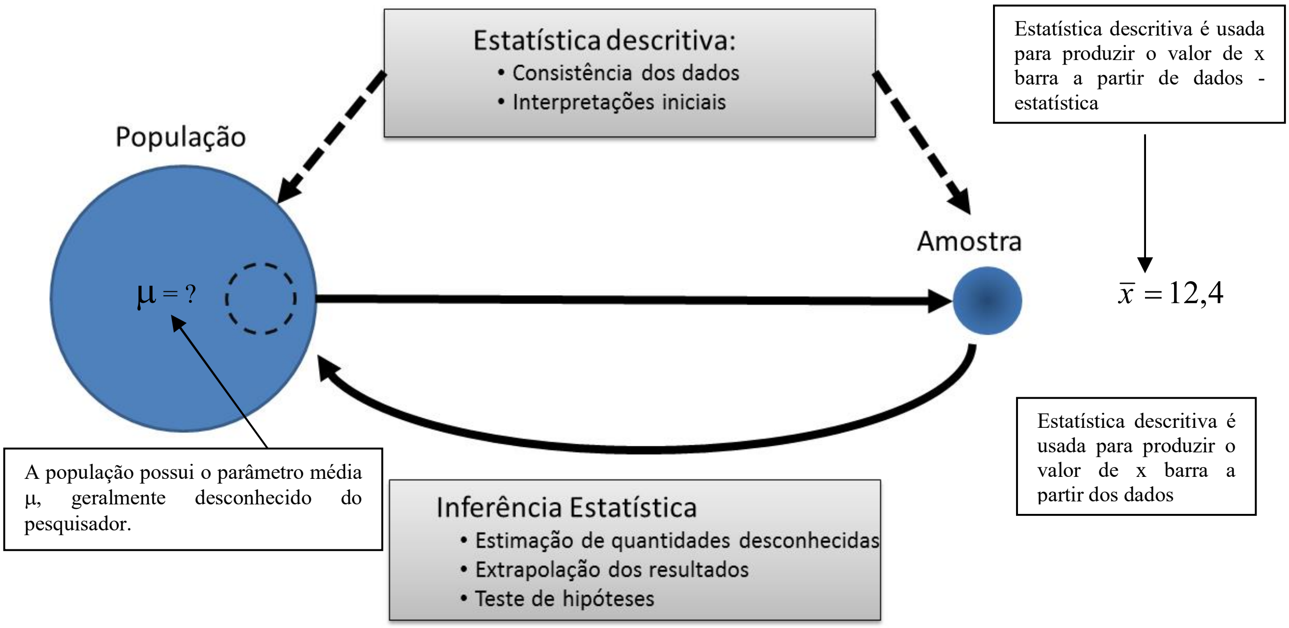
Dado: esse termo se refere ao registro das medições de características de interesse. Assim, as características tipo sanguíneo e altura de alguns, ou todos, os elementos de uma população são avaliadas e registradas. Os resultados desses processos são obtidos na forma de dados. Assim, em um ensaio experimental ou levantamento, o pesquisador terá medido, ou observado, as características que compõe a amostra e as terão registradas em forma de dados. Entretanto, o mesmo não será verdade no caso da população. Tomemos como exemplo um experimento no qual temos por objetivos realizar um teste clínico para aferição da pressão sanguínea dos alunos de uma determinada universidade. Nesse caso, será impraticável medir a pressão sanguínea de todos os alunos, mas é bastante razoável fazer medições em uma amostra de 50 dessas pressões sanguíneas.

Variável: Uma característica que pode diferir de uma entidade biológica para outra é denominada variável. É a característica de estudo do pesquisador. As informações a respeito das variáveis de interesse são armazenadas na forma de dados.

1.2.2 Parâmetros estatísticos

Os conceitos de **parâmetros** e **estatísticas** se relacionam fortemente aos conceitos de **população** e **amostra**. Um **parâmetro** é definido como qualquer resumo dos elementos de uma **população**, enquanto o resumo provável de elementos de uma **amostra** é chamado de **estatística (medida, métrica)** (não confundir com o nome da disciplina **Estatística**). Assim, a pressão sanguínea **<u>média</u>** de ***todos*** os alunos de uma universidade seria um **parâmetro** enquanto que a pressão sanguínea **<u>média</u>** dos alunos de uma determina turma (***amostra***) dessa universidade seria uma **estatística.**

Os valores dos **parâmetros** de uma população não são, normalmente, disponíveis ao pesquisador. Por outro lado, os valores das **estatísticas** estão prontamente disponíveis.



Observe que os **parâmetros** são representados por letras gregas, enquanto as **estatísticas** são representadas pelo alfabeto romano ou por uma forma dele. Por exemplo, a média de uma população é representada pela letra grega μ (pronuncia-se "mi") enquanto o mesmo resumo de dados de uma amostra é representada por $\bar{x}$ (pronuncia-se "xis barra").

Tabela. Exemplo de parâmetros e estatísticas.

Resumo	Parâmetro	Estatística
Média	μ	$\bar{x}$
Variância	σ^2	s^2
Desvio Padrão	σ	s
Correlação	ρ	r

1.3 Estatística Descritiva

1.3.1 Introdução

Em alguma fase de seu trabalho o pesquisador vê-se às voltas com o desafio de analisar e entender um conjunto de dados relevantes ao seu objeto de estudo. Se forem informações sobre uma amostra ou população, ele necessitará resumir os dados com a finalidade de que estes sejam informativos ou para compará-los com outros resultados, ou ainda para julgar sua adequação com alguma teoria. É a análise inicial que fazemos para resumir a informação a respeito do estudo.

1.3.2 Escalas de Medidas e Tipos de Variáveis

A palavra ***medir*** significa atribuir números, letras, palavras ou outro símbolo a pessoas ou coisas com o objetivo de transmitir a informação sobre as variáveis que são medidas: exemplos: atribuímos 220 mL dL^{-1} para indicar o nível de colesterol de uma pessoa; 1,80 m para indicar a altura desse mesmo indivíduo; "F" ou "M" para representar o gênero desse indivíduo. Nesse contexto, as escalas de medidas podem ser concebidas em 4 níveis diferentes, ***nominal***, ***ordinal***, ***intervalar*** e ***razões***.

Escala Nominal: é a menos sofisticada das quatro escalas. Produz classificações com base em uma avaliação qualitativa da característica sem nenhuma informação referente à quantidade ou valor. Ou seja, não existem os conceitos de "maior" ou "menor", portanto, a comparação entre os dados deve ser feita com base em "semelhante" ou "divergente".

Escala Ordinal: Semelhante à Nominal, ela classifica as pessoas ou coisas, porém tais classificações incorporam os atributos "maior que" e "menor que". Esse sistema, apesar de ordenar, não permite a indicação em termos de quanto mais ou menos. A partir dessas duas primeiras escalas de medidas, podemos definir o primeiro tipo de variável:

Variável Qualitativa: ou seja**,** é aquela que apresenta como possíveis realizações uma qualidade (ou atributo) do indivíduo pesquisado, podendo ser:

> **a) Nominal**: é aquela para a qual não existe ordenação alguma das possíveis realizações. Exemplos: sexo, grupo sanguíneo, tipo de doença, causa da morte, cor.
> **b) Ordinal**: é aquela para a qual existe certa ordem nos possíveis resultados. Exemplos: avaliação ao nascer de animais, estágio de uma doença, aparência, classe social, grau de instrução, gestão de dor (nenhuma, leve, moderada, forte).

Continuando a definição das escalas de medidas temos:

Escala Intervalar: Nessa escala acrescenta-se o atributo "quanto mais" e "quanto menos". A temperatura é um exemplo clássico. Uma leitura de 70 medida em unidades iguais a partir de um termômetro de Célsius, representa 5 unidades em graus a mais que a leitura de 65. O mesmo acontece para as leituras de 100 e 95. Essa escala tem como deficiência a falta de um ponto zero verdadeiro. Ou seja, o ponto zero na escala não representa ausência da característica. Podemos ter uma leitura de 0 ^{o}C, e não significa que não houve temperatura, pois poderíamos ter uma leitura de -10 ^{o}C no dia seguinte. Ou seja, essa escala não permite a formação de razões (quocientes) significativas, ou seja, não podemos afirmar de maneira incontestável que uma leitura de 40 ^{o}C é o dobro daquela de 20 ^{o}C. Outros exemplos, Altitude (elevação acima do nível do mar), tempo, o potencial elétrico, as direções em um plano medidas por ângulos que tem a direção zero arbitrária.

Escalas de proporcionalidade ou razões: É semelhante à escala intervalar, exceto por possuir um ponto zero verdadeiro. Considere o peso de um corpo. Não necessitamos estabelecer um ponto zero arbitrário. O peso Zero é quase um ponto de referência natural. Por esta razão, faz sentido dizermos que um animal pesa duas vezes mais que um outro, ou que seu peso aumentou 2%. O quociente entre dois valores de peso tem significado verdadeiro, por isso, chamamos está escala de escala das razões ou de proporcionalidade.

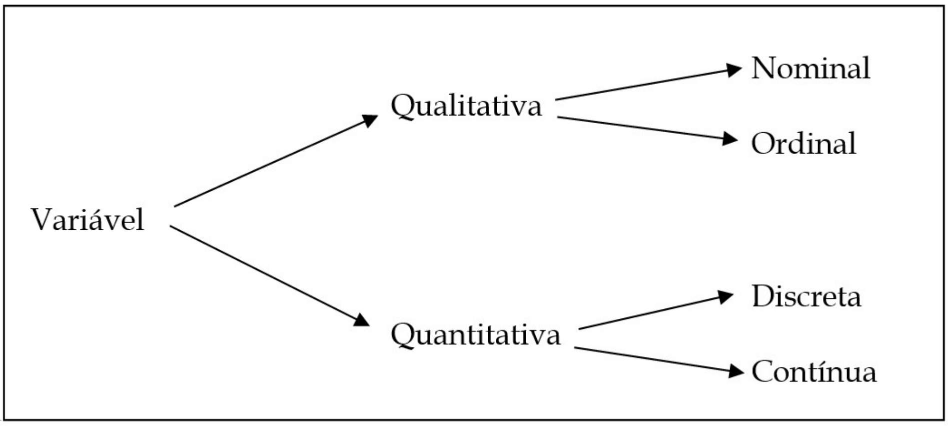
Dados Contínuos e Discretos: Existem características cujos dados podem assumir, qualquer valor em uma escala especificada. Por exemplo, uma pessoa pode pesar 70 kg e outra 71 kg. Mas é possível encontrarmos pesos entre esses dois, como 70,5 kg. Assim como é possível encontrarmos peso entre 70 e 70,5 kg, que seria 70,25 kg. Portanto, a precisão da medida dependerá da sensibilidade do instrumento utilizado para realiza-la. Esses dados são chamados de **contínuos**. Por outro lado, temos os dados **discretos**, cujos valores não existem em uma série contínua.

A partir das definições de escalas de medidas (intervalar e das razões) e dos tipos de dados (contínuos e discretos), podemos definir o segundo tipo de variável existente na **estatística:**

Variável Quantitativa é aquela que apresenta como possíveis realizações (valores) números resultantes de uma contagem ou mensuração, podendo ser:

> **a) Discreta**: é aquela cujos possíveis valores formam um conjunto finito ou enumerável de números e que resultam, frequentemente, de uma contagem e não de mensurações em uma escala contínua. Exemplos: número de filhos, número de células, número de ovos, número de ácaros ou insetos em uma planta.
>
> **b) Contínua**: é aquela cujos possíveis valores formam um intervalo de números reais e que resultam, normalmente, de uma mensuração. Exemplos: peso, altura, produção de leite, pressão arterial, teor de nitrogênio no solo ou na planta.

Em resumo, as variáveis são classificadas, em **estatística/bioestatística**, como:



1.4 Conceitos fundamentais

1.4.1 Somatório

Apesar de existir vários tipos de variáveis, é muito comum em Estatística trabalhar-se com variáveis quantitativas, que são simbolizadas por letras maiúsculas como X, Y, Z, etc. As observações ou dados, por sua vez, são representadas pelas mesmas letras minúsculas, como x, y, z, etc. Em adição, os dados são identificados por um índice, ou um contador (geralmente utilizamos as letras *i, j, k, l)* para indicar tratar da 1ª observação, 2ª observação e assim por diante. Portanto, o símbolo x_1 representa a 1ª observação do conjunto de dados referente à variável quantitativa X.

Durante os mais variados procedimentos estatísticos, é muito comum o cálculo de somas de termos, ou somas de termos ao quadrado, cálculo de médias, entre outras, então, é usual representarmos somas por um operador chamado **somatório** que é representado pela letra grega "sigma" maiúscula Σ. Assim, por exemplo, a soma de 4 elementos:

$$x_1 + x_2 + x_3 + x_4$$

É representa em notação de somatório da seguinte forma:

$$\sum_{i=1}^{4} x_i$$

ou seja, corresponde à soma dos termos x_i onde o contador i varia de 1 a 4.

O número de elementos é dado por n, nesse caso, n=4. Portanto, podemos representar a soma de todos os elementos de uma variável como:

$$\sum_{i=1}^{n} x_i$$

Em função de sua própria definição, o operador somatório possui algumas regras, dadas a seguir:

1. Se k é uma constante, e n é número de elementos, então:

$$\sum_{i=1}^{n} k = k + k + \ldots + k = nk$$

2. Se k é uma constante e x_i valores de uma variável quantitativa, então:

$$\sum_{i=1}^{n} kx_i = kx_1 + kx_2 + \ldots + kx_n = k(x_1 + x_2 + \ldots + x_n) = k\sum_{i=1}^{n} x_i$$

3. O somatório de uma soma de variáveis é igual à soma dos somatórios de cada variável.

$$\sum_{i=1}^{n}(x_i + y_i + z_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} z_i$$

4. Em consequências das regras 1, 2, e 3, se a e b são constantes, então:

$$\sum_{i=1}^{n}(a + bx_i) = \sum_{i=1}^{n} a + \sum_{i=1}^{n} bx_i = na + b\sum_{i=1}^{n} x_i$$

Exemplos

a) Expresse as seguintes somas usando a notação de somatório:

a. $y_1 + y_2 + \ldots + y_{15} = \sum_{i=1}^{15} y_i$

b. $x_1{}^2 + x_2{}^2 + \ldots + x_n{}^2 = \sum_{i=1}^{n} x_i{}^2$

c. $z_1{}^1 + z_3{}^2 + z_5{}^3 \ldots + z_{59}{}^{30} = \sum_{i=1}^{30} z_{2i-1}^{i}$

d. $\log y_1 + \log y_2 + \ldots + \log y_{12} = \sum_{i=1}^{12} \log y_i$

e. $(x_1 - 1) + (x_2^2 - 2^2)^2 + (x_3^3 - 3^3)^3 \ldots + (x_n^n - n^n)^n = \sum_{i=1}^{n}(x_i^i - i^i)^i$

b) Sabendo que:

$$\sum_{i=1}^{4} x_i = 16, \quad \sum_{i=1}^{4} x_i^{\,2} = 84, \quad \sum_{i=1}^{4} x_i^{\,3} = 496,$$

Determine o valor numérico das expressões:

a. $\sum_{i=1}^{4}\left(x_i^{\,3}-25\right)=\sum_{i=1}^{4}x_i^{\,3}-\sum_{i=1}^{4}25=496-n(25)=496-100=396$

b. $\sum_{i=1}^{4}(3x_i-15)^3$

Lembrando que:

$(a - b)^3 = a^3 - 3a^2b + 3ab^2 - b^3$

$$\sum_{i=1}^{4}(3x_i-15)^3=\sum_{i=1}^{4}\left(27x_i^3-405x_i^2+2025x_i-3375\right)=$$

$$\sum_{i=1}^{4}27x_i^3-\sum_{i=1}^{4}405x_i^2+\sum_{i=1}^{4}2025x_i-\sum_{i=1}^{4}3375=$$

$$27\sum_{i=1}^{4}x_i^3-405\sum_{i=1}^{4}x_i^2+2025\sum_{i=1}^{4}x_i-4(3375)=$$

$$(27\times 496)-(405\times 84)+(2025\times 16)-(4\times 3375)=-1728$$

1.4.2 Métodos de Numeração

Antes de iniciarmos os estudos de estatística, faz-se necessário uma pausa para relembrarmos como enumerar, ou seja, devemos estudar os procedimentos sistemáticos de contagem ou enumeração.

Regra da Multiplicação (princípio multiplicativo – regra do E): Suponha-se que um procedimento denominado 1 possa ser executado de n_1 maneiras. Admita-se que um segundo procedimento, denominado 2, possa ser executado de n_2 maneiras. Suponhamos, também, que cada maneira de executar 1 possa ser seguida por qualquer daquelas para executar 2. Então, um procedimento formado por 1 ***e*** 2 poderá ser executado de:

$$n_1 \times n_2 \text{ maneiras.}$$

Exemplo: Muitos programas de melhoramento adotam o uso de escores de avaliação visual para estimar a composição da carcaça dos animais e a rapidez com que esses chegarão ao abate, um animal que será avaliado quanto à sua Conformação, Precocidade e Musculatura, poderá receber 3 classificações para Conformação, enquanto que para Precocidade e Musculatura, esse poderá receber 4 classificações, consequentemente existem 3 . 4 . 4 = 48 maneiras que o animal pode ser classificado

Regra da Adição (princípio aditivo – regra do OU): Suponha-se que um procedimento denominado 1 possa ser executado de n_1 maneiras. Admita-se que um segundo procedimento, denominado 2, possa ser executado de n_2 maneiras. Além disso, suponha-se que não seja possível que ambos os procedimentos 1 e 2 sejam realizados em conjunto. Então, o número de maneiras pelas quais podemos realizar 1 ***ou*** 2 será:

$$n_1 + n_2 \text{ maneiras.}$$

Exemplo: suponha-se que estejamos planejando uma visita técnica ao um produtor e devemos escolher entre o transporte por ônibus, ou por trem. Só existem 3 rodovias e duas ferrovias, então existem 3 + 2 = 5 caminhos disponíveis para a viagem.

Permutações: Suponha-se que nós temos n objetos diferentes. De quantas maneiras nPn poderemos dispor (permutar) esses objetos? Por exemplo, se tivermos os objetos a, b, c, poderemos permutá-los como:

abc, acb, bac, bca, cab, cba

Ou seja, de 6 maneiras diferentes. Considera-se, em geral, o seguinte esquema: Permutar os n objetos equivale a coloca-los dentro de uma caixa com *n* compartimentos, em alguma ordenação. Dentro das caixas, apresentam-se as opções para disposição de objetos.

n	n-1	...	3	2	1
1	2	...	n-2	n-1	n

O primeiro compartimento pode ser ocupado por qualquer uma das n maneiras, o segundo compartimento por qualquer uma das (n - 1) maneiras, ..., e o último compartimento apenas por 1 maneira. Portanto, **aplicando-se a regra da multiplicação**, verificamos que a caixa poderá ser carregada de n(n-1).(n-2) ... 1 maneiras. Esse número aparece tão frequentemente em Matemática que se adotam um nome e um símbolo para ele.

Definição. Sendo n um número inteiro positivo, definimos como $n!$ = (n)(n-1)(n-2) ... 1 e o denominamos **fatorial de n**. Também definimos 0! = 1.
Assim, o número de permutação de *n* objetos diferentes é dado por:

$$nPn = n!$$

Arranjos: Considerando-se novamente o *n* objetos diferentes. Agora desejamos escolher *r* desses objetos, $0 \leq r \leq n$ e permutar os r objetos escolhidos (**ou seja, considerando a sua ordem**). Denotaremos o número de maneiras de se fazer isso (arranjos) por nPr. Recorremos novamente ao esquema anterior, de encher uma caixa com *n* compartimentos. Desta vez, simplesmente paramos depois que o compartimento *r* tenha sido ocupado.

n	n-1	...	n-(r-1)	n-r	...	3	4	1
1	2	...	r	r+1	...	n-2	n-1	n

Assim, o primeiro compartimento pode ser ocupado por n maneiras, o segundo por (n – 1) maneiras... e o de ordem r de n – (r – 1) maneiras. Portanto, o procedimento poderá ser executado aplicando-se a regra da multiplicação:

$$n(n-1)(n-2) \ldots (n-r+1)$$

Observe que as maneiras de preenchimento da caixa após a posição r, não nos interessam, então, temos que descontar do total de maneiras de ser permitas n objetos, n – r maneiras de permita-los.

nPn = n!

n	n-1	...	n-(r-1)	n-r	...	3	4	1
1	2	...	r	r+1	...	n-2	n-1	n

Permutações que estamos interessados
nPr

Permutações que não interessam
(deve-se descontar essa parte)
(n-r)!

Assim, podemos escrever o Arranjo por meio da notação fatorial definida anteriormente, ou seja:

$$n\mathrm{Pr} = \frac{n!}{(n-r)!}$$

Combinações: Considerando, novamente, n objetos diferentes. Agora trataremos da contagem do número de maneiras de escolher r dentre esses n objetos **sem considerar a sua ordem.** Por exemplo, temos os objetos a, b, c, d, para r =2; desejamos contar *ab*, *ac*, *ad*, *bc*, *bd*, *cd*; por outras palavras, não contaremos *ab* e *ba*, pois os mesmo objetos são incluídos e somente a ordem é diversa.

Para obtermos o resultado geral, recordaremos a fórmula deduzida acima: o número de maneira de escolher r objetos dentre n e permutar os r objetos é n!/(n-r)!. Assim, para definirmos a combinação desse r objetos, sem considerar a ordem, vamos defini-la como nCr. Observe que uma vez que r objetos tenham sido escolhidos, existirão r! maneiras de permutá-los. Consequentemente, aplicando-se a regra da multiplicação, temos que:

$$nCr = \frac{n!}{r!(n-r)!}$$

Este número surge em muitas passagens na Matemática e, por isso, um símbolo especial é empregado para ele. Escrevemos:

$\binom{n}{r} = \frac{n!}{r!(n-r)!}$ sendo definido para n inteiro e positivo e r inteiro tal que $0 \le r \le n$.

Exemplos:

a) Dentre 8 pessoas, quantas comissões de 3 membros podem ser escolhidas? Desde que duas comissões sejam a mesma comissão se forem construídas pelas mesmas pessoas (a ordem não importa) teremos:

$${}_8C_3 = \binom{8}{3} = \frac{8!}{3!(8-3)!} = \frac{8.7.6.5!}{3!.5!} = \frac{8.7.6}{3.2.1} = \frac{8.7.6}{6} = \frac{8.7}{1} = 56 \text{ comissões possíveis}$$

b) Com bandeiras diferentes, quantos sinais feitos com 3 bandeiras se podem obter? Apesar desse problema parecer-se muito com o anterior, a ordem de escolhas das bandeiras acarreta diferença e, por isso, temos:

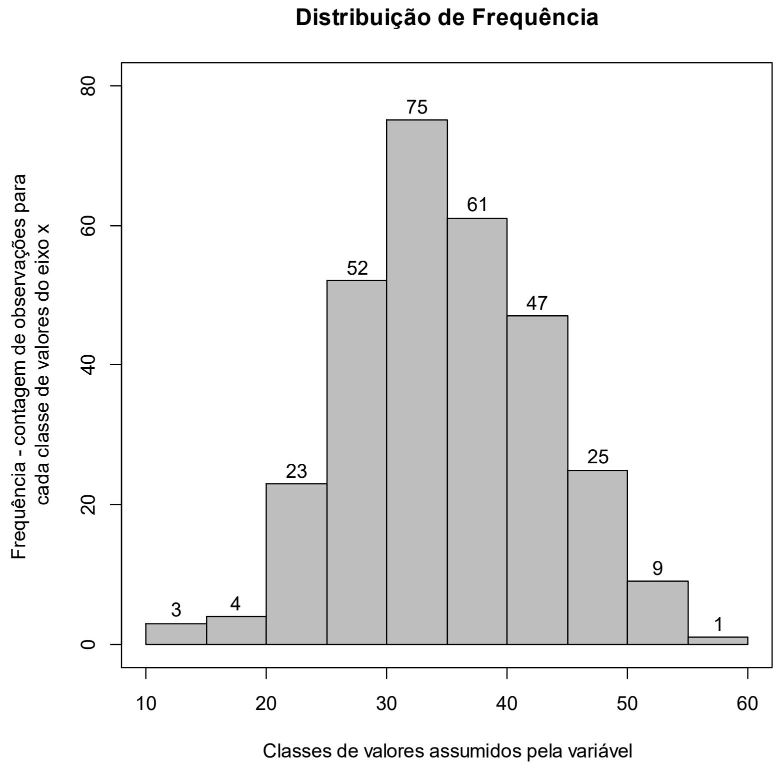
$$ {}_8P_3 = \frac{8!}{(8-3)!} = 336 \text{ sinais} $$

c) Um grupo de 8 pessoas é formado de 5 homens e 3 mulheres. Quantas comissões de três pessoas podem ser constituídas, incluindo exatamente dois homens? Aqui devemos primeiramente escolhe 2 homens entre 5 e uma mulher entre 3. Aplicando-se a regra da multiplicação.

$$ \binom{5}{2}.\binom{3}{1} = \frac{5!}{2!(5-2)!}.\frac{3!}{1!(3-1)!} = \frac{5.4.3!}{2!.3!}.\frac{3.2!}{1!.2!} = \frac{5.4}{2.1}.\frac{3}{1} = \frac{5.2}{1}.3 = 30 \text{ comissões diferentes.} $$

1.5 Distribuição de frequências de uma variável

Quando se estuda uma variável, deve-se conhecer a distribuição de frequência dessa variável por meio das possíveis realizações (dados) da mesma. Ver-se-á aqui uma maneira de disposição de um conjunto de valores, de modo a termos uma ideia global sobre estes valores, ou seja, de sua distribuição.



EXEMPLO: Um pesquisador está interessado em fazer um levantamento sobre alguns aspectos zootécnicos dos animais da Fazenda Z, ele elaborou a Tabela 1. De um modo geral, para cada elemento investigado, tem-se associado um resultado, correspondendo à realização de uma variável. Para a variável ***sexo***, por exemplo, cada animal está associado à realização "*macho*" ou "*fêmea*". Observa-se que o pesquisador colheu informações sobre seis variáveis: ***Pai***, ***Sexo***, ***Peso ao Nascer (PN)***, ***Peso aos 12 Meses de Idade (P12)***, ***Escores (Nota)*** de conformação (C), precocidade (P) e musculatura (M) aos 12 meses de idade (os escores foram obtidos utilizando-se uma escala de um a dez, sendo que as notas mais altas indicam a presença mais marcante da característica) e ***Avaliação*** ao nascer (*R* para PN ≤ 24 kg; *M* para 25 ≤ PN ≤ 29 kg; *E* para PN ≥ 30 kg).

A Tabela 2 é uma **Tabela de Frequência** para a ***variável quantitativa discreta* Nota**. As classes são representadas pelos diferentes valores que a variável assume (5, 7 e 10). No caso de uma variável qualitativa, o procedimento é o mesmo.

A **Frequência Absoluta** (n_i) é definida como o número de realizações no conjunto de dados pertencentes à classe em questão, no nosso exemplo, ocorreram 8 realizações da **Nota 5**; 32 realizações da **Nota 7** e, 10 realizações da **Nota 10**. A **Frequência Relativa ou proporção** (f_i) é definida como a proporção de cada realização em relação ao **Total** de observações.

$f_i = \frac{n_i}{n}$, onde n é o número total de observações (50 no nosso exemplo).

Tabela 1. Informações sobre o número (Nº), pai, sexo, peso ao nascer (PN), peso aos 12 meses (P12), Nota (escore) aos 12 meses de idade e Avaliação ao nascer de 50 animais da Fazenda Z (dados hipotéticos).

Nº	Pai	Sexo	PN (kg)	P12 (kg)	Nota	Avaliação
1	A	macho	22	212	5	R
2	A	fêmea	24	226	5	R
3	A	fêmea	24	196	5	R
4	A	macho	29	219	7	M
5	A	macho	27	211	7	M
6	A	macho	26	210	7	M
7	B	fêmea	20	190	5	R
8	C	macho	32	262	10	E
9	C	fêmea	27	218	7	M
10	A	macho	28	218	7	M
11	C	fêmea	28	202	7	M
12	C	fêmea	33	198	10	E
13	A	fêmea	23	138	5	R
14	C	fêmea	29	194	7	M
15	A	fêmea	21	184	5	R
16	C	fêmea	28	190	7	M
17	C	fêmea	34	215	10	E
18	C	macho	28	228	7	M
19	C	macho	28	250	7	M
20	A	macho	24	255	7	R
21	C	fêmea	31	247	10	E
22	A	fêmea	26	215	7	M
23	C	fêmea	30	244	10	E
24	B	fêmea	25	162	7	M
25	B	fêmea	27	170	7	M
26	B	fêmea	26	198	7	M
27	B	macho	30	177	10	E
28	B	fêmea	27	188	7	M
29	B	fêmea	27	136	7	M
30	C	fêmea	35	195	10	E
31	B	macho	29	246	7	M
32	C	fêmea	24	164	5	R
33	B	macho	25	192	7	M
34	A	fêmea	25	192	7	M
35	C	fêmea	25	175	7	M
36	C	macho	30	230	10	E
37	C	fêmea	27	174	7	M
38	C	fêmea	25	150	7	M

39	C	macho	27	185	7	M
40	B	macho	24	200	7	R
41	C	macho	29	183	7	M
42	C	fêmea	20	150	5	R
43	B	fêmea	26	133	7	M
44	C	fêmea	25	141	7	M
45	C	fêmea	28	162	7	M
46	C	macho	34	210	10	E
47	C	macho	28	201	7	M
48	B	fêmea	28	172	7	M
49	B	macho	35	196	10	E
50	B	macho	27	184	7	M

Tabela 2. Distribuição de frequências dos animais da Fazenda Z, segundo a **Nota** (escore) de C, P ou M aos 12 meses de idade.

Nota (x_i)	**Frequência absoluta** (n_i)	**Frequência relativa** ($f_i=n_i/n$)	**Porcentagem** ($100 \times f_i$)
5	8	0,16	16
7	32	0,64	64
10	10	0,20	20
Total(n)	**50**	**1,00**	**100**

A Tabela 3 é a tabela de frequência para uma ***variável quantitativa contínua*** **P12** e, nesse caso, as classes são intervalos reais (k). Inicialmente, devemos fixar o número de intervalos, a regra geral em diz que: uma boa representação apresenta um número de intervalos ***nunca inferiores a 5 ou superiores a 15***, pois com um pequeno número de classes, perde-se informação, e com um grande número de classes, o objetivo de resumir os dados fica prejudicado. Para exemplificar, vamos fixar o número de intervalos em 5 (k = 5). Tais intervalos são subintervalos da **Amplitude Total (Δ)** dos dados, ou seja, diferença entre a ***maior*** e a ***menor*** observação, correspondendo o intervalo de valores numéricos que contém todos os dados observados.

Tabela 3. Distribuição de frequências dos animais da Fazenda Z, por classe de **P12 (**pesos aos 12 meses - kg).

Classes de Pesos (k)	n_i	pmc_i	f_i	%	d_i	N_i	F_i (N_i/n)	$100 \times F_i$
133 |---- 159	6	146	0,12	12	0,0046	6	0,12	12
159 |---- 185	11	172	0,22	22	0,0085	17	0,34	34
185 |---- 211	17	198	0,34	34	0,0131	34	0,68	68
211 |---- 237	10	224	0,20	20	0,0077	44	0,88	88
237 |---- 263	6	250	0,12	12	0,0046	50	1,00	100
Total (n)	**50**	**-**	**1,00**	**100**	**-**	**-**	**-**	**-**

Fonte : Tabela 1; N_i = frequência acumulada até a i-ésima classe; F_i = frequência acumulada relativa; 100 × F_i = porcentagem acumulada.

Amplitude Total (Δ): Para a variável ***Peso aos 12 meses* (P12)**, temos:

$$\Delta = Máximo - Mínimo$$

$$\Delta = 262 - 133$$

$$\Delta = 129$$

A **Amplitude de classe** (subintervalo, denominado Δ_i) é determinado dividindo-se a **Amplitude Total** em um número conveniente de subintervalos que tenham a mesma amplitude. Isto é feito dividindo-se a amplitude total pelo número de classes desejável (k=5 no nosso exemplo). Pode-se arredondar esse quociente para um número exato de subintervalos, acrescentando-se ao conjunto de dados, valores com frequência nula.

Amplitude de classe: Δ_i = amplitude do intervalo da classe i;

$$\Delta_i = \frac{\Delta}{k}$$

$$\Delta_i = \frac{129}{5} = 25{,}8$$

$$\Delta_i \cong 26$$

OBS: Uma forma de determinação de um número razoável, k de classes consiste em aplicar a fórmula de **Sturges**, que sugere o cálculo de k mediante a expressão:

$$k = 1 + \log_2 n$$

$$ou\ seja:$$

$$k = 1 + \frac{\log n}{\log 2}$$

Por exemplo, para $\boldsymbol{n}$ = 50:

$$k = 1 + log_2 50$$
$$k = 1 + \frac{log\ 50}{log\ 2}$$
$$k = 1 + \frac{3{,}91}{0{,}69}$$
$$k = 1 + 5{,}7 \cong 7$$

Em caso de uma quantidade muito grande de dados quantitativos discretos, ou seja, de valores que a variável assume, é conveniente construir a tabela de frequências do mesmo modo que é feito para uma variável contínua, isto é, considerando classes como subintervalos.

Como visto, a amplitude do intervalo de classe (Δ_i) na **Tabela 3** foi determinada dividindo-se a amplitude total (Δ) pelo número de classes desejável (k = 5). Observe que o limite superior da última classe foi o valor 263, com frequência nula no nosso conjunto de dados, ou seja, o valor 263 não foi observado. Tal procedimento garante que o valor máximo do conjunto de dados seja incluído na última classe. Portanto, o símbolo adotado (|----), significa que o extremo inferior da classe está incluído nela e o extremo superior excluído. Assim, o valor 159, por exemplo, está incluído na segunda classe. Pode-se usar também nas classes a notação [;), cujo significado é o mesmo do anterior, ou seja, fechado à esquerda e aberto à direita.

Procedendo-se como na Tabela 3, ao resumir os dados referentes a uma variável quantitativa contínua, ***perde-se alguma informação***. Por exemplo, não se tem informação de como se distribuem os 6 pesos na primeira classe, a não ser que se investigue os dados originais (Tabela 1). Sem perda de muita precisão, pode-se supor que todos os pesos de uma determinada classe sejam iguais ao ponto médio dessa classe (pmc_i), isto é, no caso da primeira, 146 kg.

Ponto médio da classe *i* (pmc_i): é definido como o valor médio entre os limites superiores e inferiores de uma determinada classe (i).

$$pmc_i = \frac{(LI_i + LS_i)}{2},$$

Assim, para as classes da Tabela 3, temos:

$$pmc_1 = \frac{(133 + 159)}{2} = 146$$
$$pmc_2 = \frac{(159 + 185)}{2} = 172$$
...
$$pmc_5 = \frac{(237 + 263)}{2} = 250$$

Densidade de frequência ou simplesmente densidade (*d_i*): definida como o quociente entre a área pela amplitude de classe, utilizada na construção do gráfico histograma, que faz com que esse não fique distorcido quando se utiliza amplitude de classes diferentes. Para que a área do retângulo de uma respectiva classe no histograma se proporcional à f_i, a sua altura deve ser proporcional a f_i/Δ_i, que é denominada ***densidade de frequência da i-ésima classe.***

$$d_i = \frac{f_i}{\Delta_i}$$

1.6 Representação gráfica da distribuição de frequências

Gráfico é uma apresentação de dados estatísticos na forma visual. Sua importância é consagrada em todas as ciências, pois é a maneira mais simples de resumir e apresentar a informação. Os principais tipos de gráficos usados na representação estatística são:

a. **Gráfico em barras**: é um tipo de gráfico que se obtém locando os valores no eixo horizontal e traçando-se em cada um deles um segmento vertical de altura proporcional à respectiva frequência (relativa ou absoluta). Esse tipo de gráfico se adapta melhor às ***variáveis quantitativas discretas*** ou ***qualitativas ordinais***.

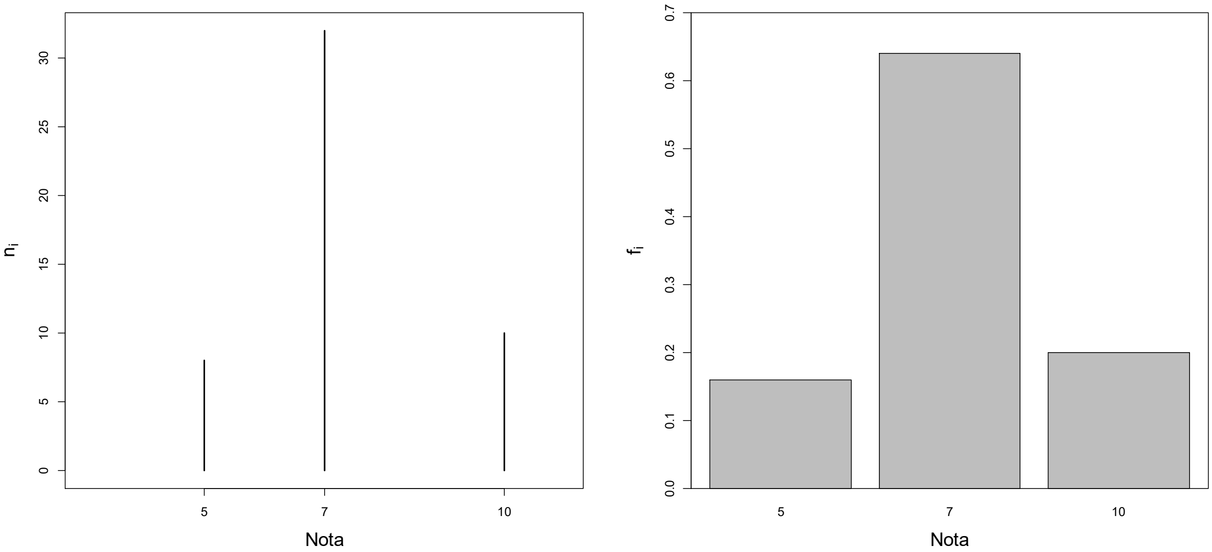


Gráfico 1. Duas representações de gráficos de barras dos dados da Tabela 2.

b. **Histograma**: é um conjunto de retângulos, com bases sobre um eixo horizontal, divididos de acordo com os tamanhos das classes (Δ_i), com centros nos pontos médios das classes (pmc_i) e áreas proporcionais às ***frequências*** (f_i ou n_i). Em certos casos, é interessante que a área total da figura seja igual a 1, correspondendo à soma total das proporções (f_i). Então, para construção do histograma, sugere-se usar no eixo das ordenadas os valores de f_i / Δ_i (***densidade de frequência***), ou seja, da medida que indica qual a concentração por unidade da variável.

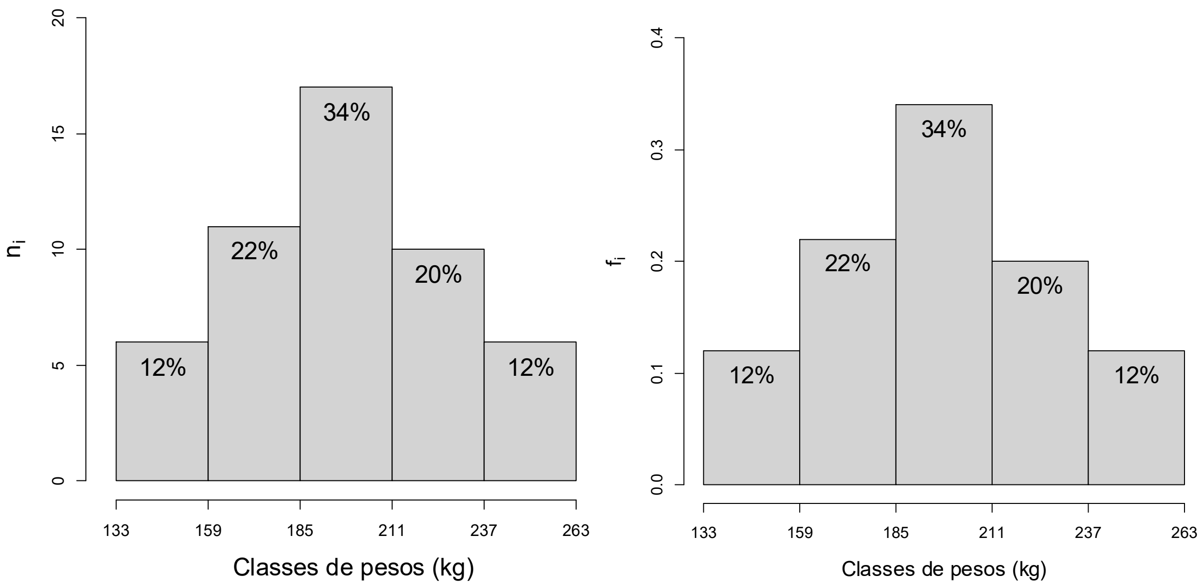


Gráfico 2. Histograma da variável peso aos 12 meses (Tabela 3), utilizando a frequência absoluta ou relativa.

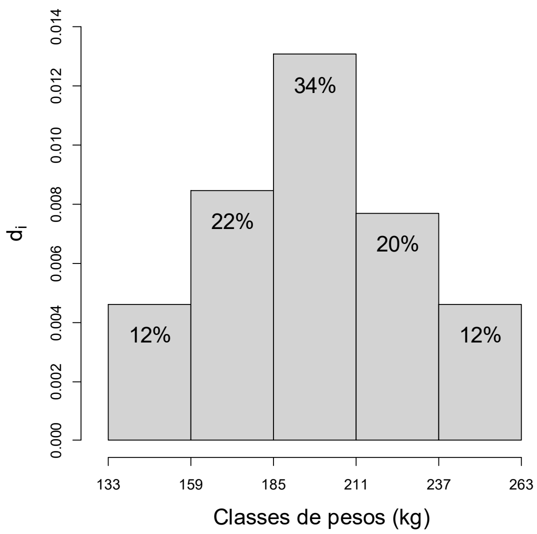


Gráfico 3. Histograma da variável peso aos 12 meses (Tabela 3), utilizando a densidade de proporção.

c. **Polígono de frequências**: é um gráfico que se obtém unindo por uma poligonal os pontos correspondentes às ***frequências***, das diversas classes, centradas nos respectivos pontos médios. Para se obter as interseções do polígono com o eixo horizontal, cria-se em cada extremo do histograma uma classe com frequência nula.

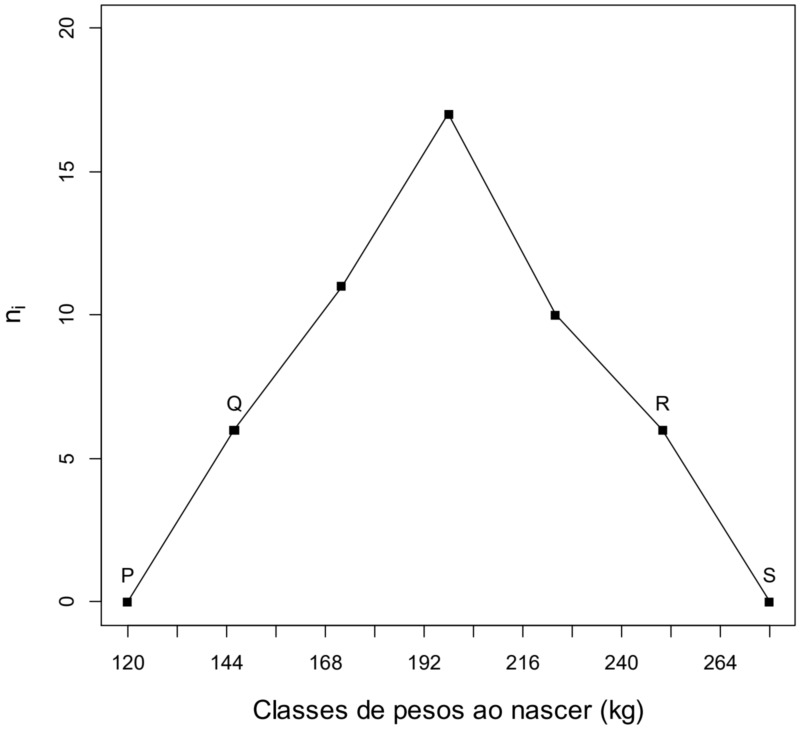


Gráfico 4. Polígono de frequência para os dados da Tabela 3. Note que ao construir o polígono de frequência foram acrescentados os segmentos PQ e RS, que vão ter ao ponto médio imediatamente inferior e superior e cujas frequências são nulas. Nesse caso, a soma das áreas dos retângulos do histograma é igual área total limitada pelo polígono de frequência e o eixo horizontal.

d. **Polígono de frequências acumuladas percentuais** (ou ogiva percentual): é um gráfico poligonal ascendente que representa a ***frequência acumulada*** abaixo de qualquer limite superior de classe. No eixo horizontal colocam-se as extremidades de classe, e no eixo vertical, as frequências acumuladas percentuais.

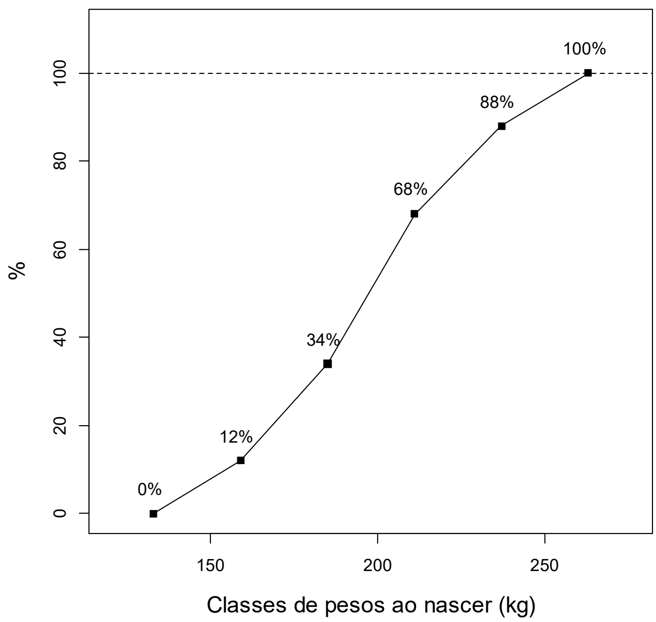


Gráfico 5. Polígono de frequência acumulada percentual (ou ogiva percentual) dos dados da Tabela 3.

e. **Gráfico em linha**: é um dos mais importantes gráficos; representa observações feitas ao longo do tempo, em intervalos iguais ou não, traduzindo o comportamento de um fenômeno em certo intervalo de tempo. É bastante utilizado para mostrar tendência.

f. **Gráfico em setores**: aplicável quando as categorias (classes) básicas são quantificáveis. Toma-se um círculo (360 graus), que se divide em setores com áreas proporcionais às frequências das diversas categorias. Esse tipo de gráfico se adapta muito bem às ***variáveis qualitativas nominais***.

Exemplo. Considerando seguintes participações no mercado da venda de sêmen das raças leiteiras nacionais:

Holandês	50%	(180 graus)
Gir leiteiro	29%	(104 graus)
Jersey	10%	(36 graus)
Suíça	7%	(25 graus)
Outras	4%	(15 graus)

Observe-se que 180 graus representam precisamente 50% de 360 graus, e assim por diante.
Solução:

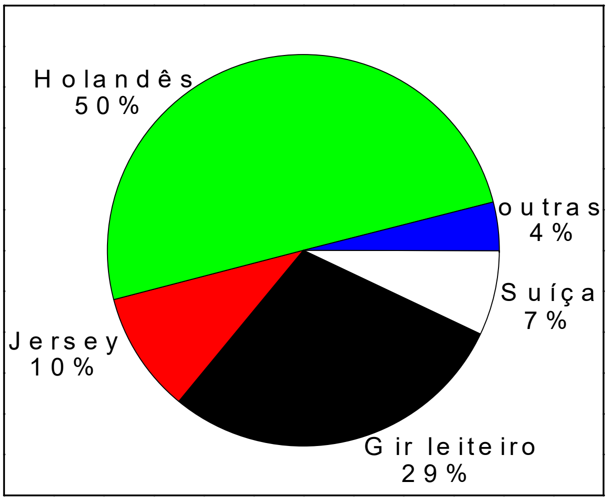


Gráfico 6. Gráfico em setores do exemplo.

Intervalos de classes desiguais

Como mencionado anteriormente, quando os comprimentos Δ_i das classes são diferentes, deve-se usar para a construção do histograma f_i/Δ_i no eixo vertical, cujos valores são muito mais informativos para compreender a distribuição, do que as frequências simplesmente. É o caso do exemplo a seguir (Tabela 4). Uma outra vantagem diz respeito à relação entre histograma e gráfico da função densidade de probabilidade, que será visto mais adiante.

Tabela 4. Distribuição de frequências dos animais da Fazenda Z, por classe de pesos ao nascer (kg).

Classes de pesos	n_i	f_i	Δ_i	f_i/Δ_i
20 |--- 23	4	0,08	3	0,0267
23 |--- 26	12	0,24	3	0,0800
26 |--- 29	20	0,40	3	0,1333
29 |--- 31	7	0,14	2	0,0700
31 |--- 37	7	0,14	6	0,0233
Total	**50**	**1,00**	**-**	**-**

Fonte : Tabela 1

f_i / Δ_i = densidade de frequência da classe *i*

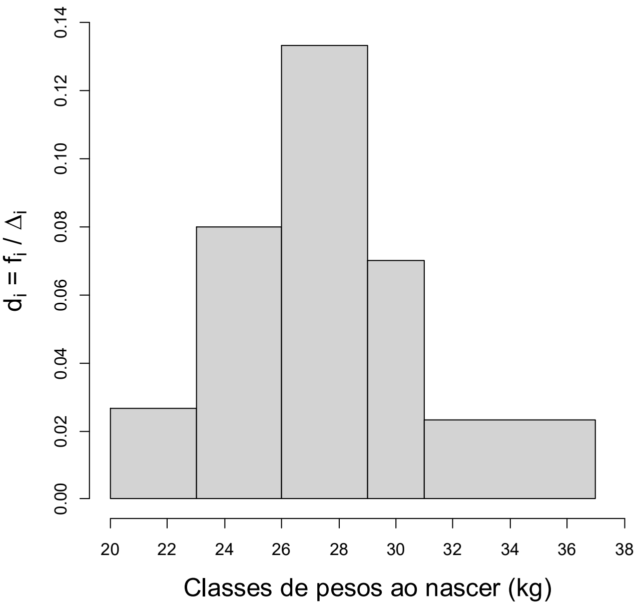


Gráfico 7. Histograma da variável peso ao nascer (Tabela 4).

Histograma para variável discreta. Do mesmo modo que usamos um artifício para representar a variável contínua como discreta, podemos usar um artifício para construir um histograma para variáveis discretas. O Gráfico 6 é um exemplo de como fica o histograma da variável nota de C, P ou M aos 12 meses de idade, segundo dados da Tabela 2.

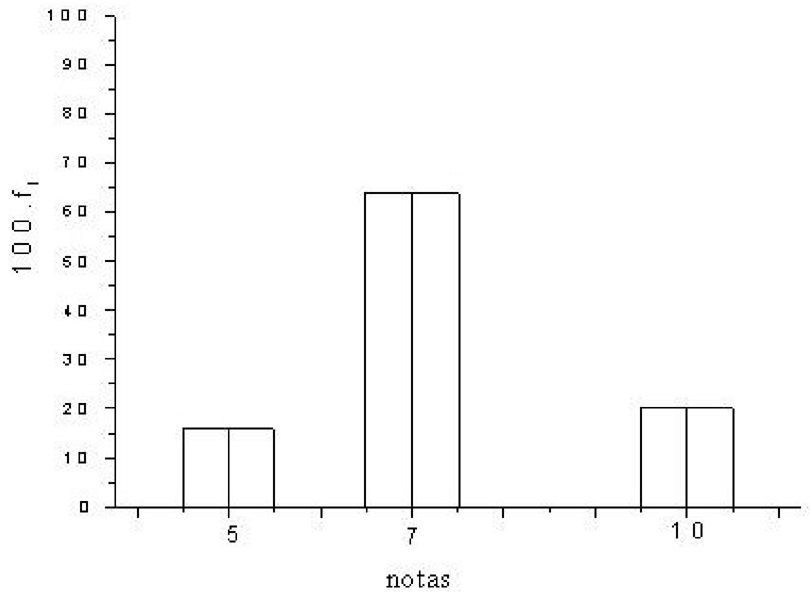


Gráfico 8. Histograma ajustado para a variável nota de C, P ou M (Tabela 2).
Note que ao construir o histograma, os centros dos retângulos foram determinados pelos valores das notas, tal que a largura de cada retângulo seja igual a um (1).

2 MEDIDAS ESTATÍSTICAS ASSOCIADAS A VARIÁVEIS QUANTITATIVAS

O resumo dos dados por meio de tabelas de frequências e gráficos de dispersão fornecem muito mais informação sobre o comportamento dos dados de uma variável do que a própria tabela original de dados. Entretanto, é necessário resumir ainda mais estes dados, apresentando alguns valores representativos da série inteira. Assim, o objetivo deve ser a caracterização do conjunto de dados por meio de medidas que resumam a informação, representando a tendência central, ou a maneira pela qual estes dados estão dispersos.

2.1 Medidas de posição ou de tendência central

Mostram o valor representativo em torno do qual os dados se distribuem. São utilizadas para sintetizar, em um único número, o conjunto de dados observados. Talvez a medida mais conhecida desse tipo seja o que normalmente é conhecido como "média" ou, mais precisamente *média aritmética* de um conjunto de dados. A média é considera a medida de posição mais importante. Podemos ter 4 tipos de médias:

Média Aritmética
Média Ponderada
Média Geométrica
Média Harmônica

2.1.1 Média Aritmética $(\bar{x})$

É a mais utilizada das medidas de posição. A média aritmética (ou simplesmente média) de um conjunto de *n* observações, x_1, x_2, ... , x_n, da variável X, é o quociente da divisão da *soma* dos valores das observações dessa variável por *n*. A média para uma amostra pode ser representada por $\bar{x}$ (xis barra) ou $\hat{m}$ (m chapéu, onde "^" denota *estimativa*). Pode-se escrever:

$$\bar{x} = \frac{x_1 + x_2 + \Lambda + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad i = 1, 2, \Lambda\ , n$$

OBS: Cada medida no conjunto de observações é referida como um valor x_i, tal que o primeiro valor é referido como x_1, o segundo como x_2, e assim por diante. O subscrito *i*, que pode ser qualquer número inteiro entre 1 e o total de valores *n*, corresponde, então, à posição de cada valor no conjunto de observações.
Para a população a média é definida como:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N} \qquad i = 1, 2, \Lambda\ , N$$

Sabemos que n_i representa a frequência absoluta de uma observação x_i, com i = 1, 2, ..., *k*, então.

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \Lambda\ + n_kx_k}{n} = \frac{\sum_{i=1}^{k} n_i x_i}{n}$$

em que $n = \sum_{i=1}^{k} n_i$; e se $f_i = \frac{n_i}{n}$ representa a frequência relativa da observação x_i, i = 1, 2, ..., k, então (1) também pode ser escrita como:

$$\bar{x} = \sum_{i=1}^{k} f_i x_i$$

Exemplo 1. Considerando as notas de C, P ou M aos 12 meses de idade dos 50 animais, apresentadas na Tabela 1,

$$\bar{x} = \frac{5+5+\cdots+7}{50} = 7{,}28$$

Usando agora a tabela de distribuição de frequência da variável Nota (Tabela 2 – Aula 1), isto é:

x_i	**5**	**7**	**10**
n_i	8	32	10
f_i	0,16	0,64	0,20

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \frac{(8\times 5 + 32\times 7 + 10\times 10)}{50} = 7{,}28 \quad \text{ou}$$

$$\bar{x} = \sum_{i=1}^{k} f_i x_i = 0{,}16\times 5 + 0{,}64\times 7 + 0{,}20\times 10 = 7{,}28$$

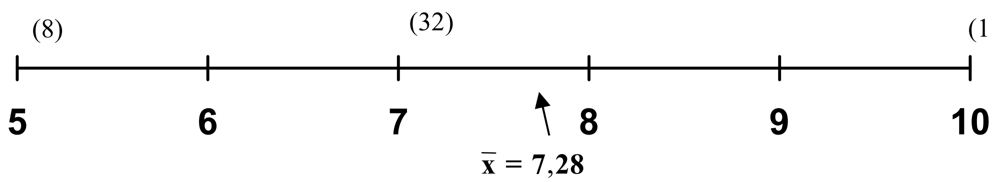


Figura 1. Média como ponto de equilíbrio, ou centro, da configuração.

2.1.2 Média ponderada $(\bar{x}_p)$

Em algumas situações, a média aritmética não é recomendada, uma vez que as observações têm graus de importância diferentes. Usa-se então a média ponderada. Chama-se média ponderada entre n observações, x_1, x_2, ..., x_n, o número:

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

onde λ_i é o peso associado à i-ésima observação (isto é, ele mede a importância relativa da i-ésima observação em relação às demais).

A média aritmética pode ser interpretada como uma média ponderada em que os pesos são todos iguais.

Exemplo 1. Calcular a média final (ponderada) na disciplina de Bioestatística, considerando que:

	Peso (λ_i)	Nota (x_i)
1ª Prova	4	6,0
2ª Prova	5	5,0
Trabalho	1	8,0

A média final é:

$$\bar{x}_p = \frac{\sum_{i=1}^{n} \lambda_i x_i}{\sum_{i=1}^{n} \lambda_i} = \frac{(4\times 6{,}0)+(5\times 5{,}0)+(1\times 8{,}0)}{4+5+1} = 5{,}7$$

2.1.3 Média Geométrica $(\bar{x}_g)$

Em casos raros, utiliza-se a média geométrica, que consiste em determinar a raiz n-ésima do produto dos *n* dados considerados.

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Exemplo: Uma represa foi infestada por uma vegetação daninha aquática a qual cobriu 12 km² da represa, com um aumento anual foi de 50%. Os dados de área da represa coberta por essa vegetação estão expressos na tabela abaixo.

Ano	1	2	3	4	5	6	7	8	9
Área (km²)	12	18	27	40,5	60,75	91,125	136,6875	205,03125	307,546875

Assim, temos um problema de porcentagem, ou seja, a cada ano a área coberta pela vegetação daninha aumenta em 50% do seu tamanho (18=12×1,5; 27=18×1,5 e assim sucessivamente). Nesse caso a média aritmética perde seu significado biológico. Observando o conjunto de dados, podemos entender que o valor de 60,75 km² (quinto período de ano) é a própria média para esse conjunto de dados. Assim, compare os valores de média aritmética e média geométrica para esse conjunto de dados:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{898{,}641}{9} = 99{,}849 \ km^2$$

O valor encontrado é bastante diferente de 60,75 km², entretanto, se utilizarmos a média geométrica:

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[9]{1,12698 \times 10^{16}} = 60,75 \text{ km}^2$$

OBS: Aplicando as propriedades dos logaritmos, também podemos escrever a média geométrica como:

$$\bar{x}_g = e^{\frac{\sum_{i=1}^{n} \ln(x_i)}{n}}$$

Esta fórmula é computacionalmente mais interessante, de fácil programação que a anterior, uma vez que não precisamos multiplicar os dados, ao invés disso, trabalhamos com a média aritmética do logaritmo natural das observações.

2.1.4 Média harmônica $(\tilde{x})$

A média harmônica é definida como o inverso da média dos inversos, ou seja:

$$\tilde{x} = \frac{1}{\left(\sum_{i=1}^{n} 1/x_i \right)/n}$$

Considere o seguinte exemplo. Um elefante possui um território o qual é um quadrado de lado igual a 2 km. Em cada manhã, o elefante anda sobre os limites de seu território (perímetro do quadrado). No início do dia ele anda o primeiro lado de seu território na velocidade de 1 km h^{-1}. Ao percorrer o segundo lado, ele aumenta a sua velocidade para 2 km h^{-1}. No terceiro lado o elefante acelera para incríveis 4 km h^{-1}, entretanto, tal esforço desgasta o animal e ele se vê forçado a percorrer o quarto e último lado de seu território com a velocidade lenta de 1 km h^{-1}. Pergunta-se, qual a velocidade média do elefante ao longo de todo o percurso?

Mais uma vez, a média aritmética não tem sentido nesse exemplo, pois, poderíamos pensar que a sua velocidade média foi:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+2+4+1}{4} = \frac{8}{4} = 2 \text{ km h}^{-1}$$

Entretanto, devemos lembrar que a velocidade média é dada por:

$$Vm = \frac{\Delta s}{\Delta t}$$

Assim, para calcular a velocidade média do elefante precisamos da distância total percorrida pelo elefante (4 × 2 = 8 km) dividida pelo tempo total gasto pelo animal. Assim, os tempos totais podem ser calculados como:

$$\Delta t = \frac{\Delta s}{Vm}$$

Lado 1 : $\Delta t = 2/1 = 2$ h

Lado 2 : $\Delta t = 2/2 = 1$ h

Lado 3 : $\Delta t = 2/4 = 0,5$ h

Lado 4 : $\Delta t = 2/1 = 2$ h

Assim,

$$Vm = \frac{8}{2 + 1 + 0,5 + 1} = 1,454 \text{ km h}^{-1}$$

Utilizando a média harmônica teríamos:

$$\tilde{x} = \frac{1}{\left(\sum_{i=1}^{n} 1/x_i\right)/n} = \frac{1}{\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{4} + \frac{1}{1}\right)/4} = 1,454$$

OBS: Para a aplicação da média harmônica, todos os dados devem ser diferentes de Zero.

2.1.5 Mediana *(Md)*

É a realização que ocupa a posição central de uma série (*n*) de observações, quando estão ordenadas em ordem crescente (Rol), nem sempre pertence ao conjunto de dados. Se *n* é ímpar, esse valor é único. Se *n* é par, *Md* é a média dos dois valores centrais.

Exemplo. Se x_i = 3, 4, 7, 8, 8 $\Rightarrow$ *Md* = 7

Acrescentando-se o valor 9 ao conjunto de valores, x_i = 3, 4, 7, 8, 8, 9

$$Md = \frac{7 + 8}{2} = 7,5$$

Assim, uma vez que o conjunto de dados está ordenado, temos a mediana como:

$$Md = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ é ímpar;} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ é par.} \end{cases}$$

Quando uma medida de posição for pouco afetada por mudanças de uma pequena porção de suas observações (dados), é dito que ela é **resistente**. A mediana é uma medida resistente, enquanto que a média não o é.
Como ilustração, tomemos as observações (dados): x_i = 5, 7, 8, 10, 12, onde

$$\bar{x} = 8,4 \text{ e } Md = 8,0$$

Substituindo, agora, o valor 12 por 120 os dados ficarão x_i = 5, 7, 8, 10, 120 e obteremos:

$$\bar{x} = 30 \text{ e } Md = 8,0$$

ou seja, a mediana não se altera enquanto a média aumentou mais de três vezes. Portanto, a mediana não é afetada por observações muito grandes ou muito pequenas, enquanto que a presença de tais extremos tem um significante efeito sobre a média. Mais adiantem estudaremos que distribuições extremamente assimétricas, a mediana é, provavelmente, uma medida de centro mais adequada do que a média. Caso contrário, a média é preferida e mais amplamente usada, isto porque a mediana carece de algumas vantagens teóricas relacionadas à inferência estatística.

2.1.6 Moda *(Mo)*

É definida como a realização mais frequente do conjunto de valores observados.
Exemplo 5. Considerando a variável nota ao nascer resumida na Tabela 2,

$Mo = 7$

Em alguns casos, a distribuição de valores pode ser amodal (não apresenta moda) unimodal (apresenta apenas um valor de moda), bimodal, trimodal, etc.
No caso de dados agrupados, definimos a moda como o **ponto médio** da classe de maior frequência (classe modal), desde que todas as classes tenham a mesma amplitude.

Exemplo. $Mo = 198$ kg para os dados da Tabela 3 do capítulo anterior.

OBS: Observe que para calcular a moda de uma variável, precisamos apenas da distribuição de frequência (contagem). Já para a mediana necessitamos minimamente ordenar as realizações da variável. Finalmente, a média só pode ser calculada para variáveis quantitativas. Portanto, essas condições limitam bastante o cálculo de medidas-resumo para as variáveis qualitativas. Para as variáveis nominais somente podemos trabalhar com a moda. Para as variáveis ordinais, além da moda, podemos usar a mediana.

2.1.7 Quantis

Se o número de observações é grande (maior do que 30) é útil estender a noção de mediana e dividir o conjunto de dados em quantis.
O quantil de ordem $100p$ de um conjunto de valores dispostos em ordem crescente é um valor tal que até ele (inclusive) haja pelo menos $100p\%$ das observações e, a partir dele (inclusive) haja pelo menos $100(1 - p)\%$ das observações $(0 < p < 1)$.
Os quantis de ordem 25, 50, 75 são chamados **quartis** (Q_1, Q_2, Q_3). Naturalmente, $Q_2 = Md$.
Os **decis** são os quantis de ordem 10, 20, ..., 90 ($D_1, D_2, ..., D_9$) e os **percentis** são os quantis de ordem 1, 2, ..., 99 ($P_1, P_2, ..., P_{99}$).

Será adotada a convenção de se tomar um valor observado para o quantil, exceto quando valores adjacentes satisfazem a definição, sendo que neste caso o quantil será tomado como a média desses valores. Isto coincide com o modo com que a mediana foi definida quando o número de observações é par. Ilustraremos, a seguir, um método para se determinar quartis, com um exemplo envolvendo poucas observações.

Exemplo. Considerando o conjunto de valores, já ordenados do menor para o maior: 93,9; 105,8; 106,5; 116,6; 125,0; 128,3; 132,1; 136,7; 152,4, obter os quartis.

Solução. O número de observações ≤ Q_1 é 0,25×9 = 2,25, ou seja 3, e ≥ Q_1 é 0,75×9 = 6,75, ou seja 7. Contando 3 valores do menor para o maior e 7 valores do maior para o menor, encontramos 106,5 e este é o valor de Q_1. Assim procedendo,

$Q_1 = 106{,}5$ *e* $Q_2 = Md = 125{,}0$ *e* $Q_3 = 132{,}1$

Acrescentando-se o valor 153,0 ao conjunto de valores, isto é 93,9; 105,8; 106,5; 116,6; 125,0; 128,3; 132,1; 136,7; 152,4; 153,0, então:

$Q_1 = 106{,}5$ $\quad Q_2 = \dfrac{125{,}0 + 128{,}3}{2} = 126{,}65$ $\quad Q_3 = 136{,}7$

2.1.8 **Média e mediana de dados agrupados**

Sempre que possível, as medidas estatísticas devem ser calculadas antes do agrupamento de dados. Não raro, entretanto, é conhecermos só o quadro de distribuição de frequência para os dados agrupados. Com os dados agrupados em classes, como já mencionado, perde-se informação sobre cada observação individual, e uma boa aproximação é supor que todos os dados, dentro de uma classe tenham seus valores iguais ao ponto médio dessa classe. Fazendo, então, $pmc_1, pmc_2, ..., pmc_k$ os pontos médios das k classes, e $n_1, n_2, . . ., n_k$ (ou $f_1, f_2, . . ., f_k$) as respectivas frequências, a média é, então, calculada como em (1) ou (2).

Exemplo 4. Considerando os dados de peso aos 12 meses agrupados em intervalos de classes (Tabela 3).

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \times pmc_i}{n} = \frac{6 \times 146 + 11 \times 172 + \Lambda\ + 6 \times 250}{50} =$$

$$\bar{x} = \sum_{i=1}^{k} f_i \times pmc_i = 0{,}12 \times 146 + 0{,}22 \times 172 + ... + 0{,}12 \times 250 = 197{,}48\ kg$$

Md = 198,0 kg = ponto médio da classe que contém a mediana (critério aproximado).

Obs. Usando os dados da Tabela 1 da aula anterior, obtemos os seguintes valores: $\bar{x} = 195{,}76\ kg\ \ e\ \ Md = 195{,}5\ kg$.

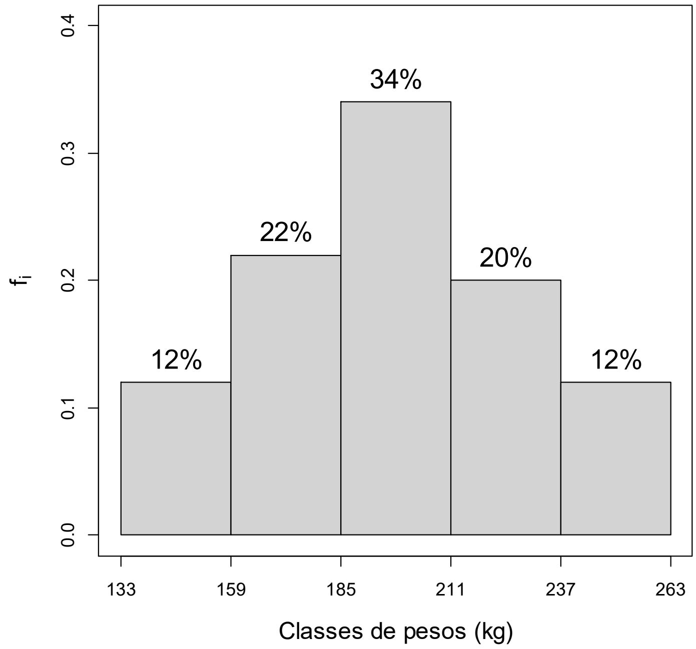
2.1.9 Quantis de dados agrupados

Processo gráfico

Histograma

Usando-se o histograma, pode ser formulado o seguinte procedimento para se encontrar quantis de uma variável com dados agrupados.
O cálculo do quantil desejado, por exemplo a mediana (2º quartil), é feito, conforme sua definição, localizando-se o ponto das abcissas que divide a área do histograma em duas partes iguais (50% para cada lado). Então, usando argumentos geométricos pode-se encontrar um ponto satisfazendo esta propriedade.

Vejamos por meio do histograma apresentado a seguir:



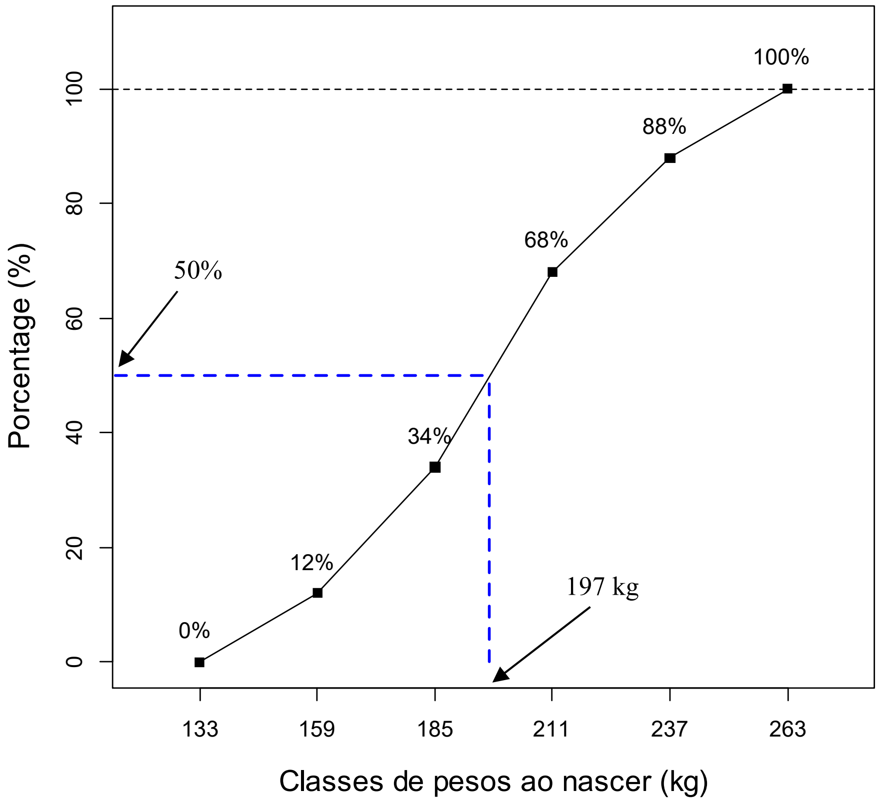
Histograma da variável peso aos 12 meses (Tabela 3)

onde a mediana irá corresponder ao valor (*Md*) no terceiro retângulo, tal que a área do retângulo de base [185, Md) e de mesma altura que o de base [185, 211) seja 16% (12% do 1º retângulo, mais 22% do 2º e 16%, de um total de 34%, do 3º, perfaz os 50%). Por meio da proporcionalidade entre a área e a base do retângulo, têm-se $\frac{211-185}{0,34} = \frac{\text{Md}-185}{0,16}$. Logo: Md = 197,24kg.

Esse procedimento de cálculo pressupõe que as observações estejam em ordem crescente e igualmente espaçadas dentro de cada classe. O cálculo dos demais quantis pode ser feito de modo análogo, ou seja, por interpolação linear, que se reduz a uma regra de três simples.

No caso de dados agrupados, outro processo gráfico bastante prático para determinação de quantis, de qualquer ordem, utiliza a ogiva percentual.

(b) Ogiva percentual



Ogiva percentual da variável peso aos 12 meses (Tabela 3)

Por este processo gráfico, de acordo com a frequência desejada (quartil, decil, percentil), traça-se uma paralela ao eixo horizontal. A partir do ponto em que esta paralela encontra a ogiva percentual, traça-se uma perpendicular ao eixo horizontal. O ponto de encontro com este eixo é o valor do quantil procurado.

2.2 Medidas de dispersão ou variabilidade

O resumo de um conjunto de dados, por meio de uma única medida representativa de posição central, esconde toda informação sobre a variabilidade do conjunto de valores. As medidas de variação medem o grau com que os dados tendem a se distribuir em torno de um valor central que, geralmente, é a média aritmética. Portanto, as tendências centrais podem não ser suficientes na descrição e discriminação de diferentes conjuntos de dados.
Exemplo. Consideremos os conjuntos de observações

A = {25, 28, 31, 34, 37}   B = {17, 23, 30, 39, 46}

Verifica-se que ambos têm a mesma média, $\bar{x}(A) = \bar{x}(B) = 31$. A identificação de cada um desses conjuntos de dados pelas suas médias, nada informa sobre as diferentes variabilidades dos mesmos. Então, é conveniente criar uma medida que sintetize a variabilidade de uma série de valores e que nos permita comparar conjuntos diferentes de valores, como os acima, segundo algum critério estabelecido.

O critério frequentemente usado para resumir a variabilidade de uma série de valores é medir a concentração dos dados em torno de sua média e a medida mais usada é a **variância**.

O princípio básico é analisar os desvios (x_i - $\bar{x}$). Assim, poderíamos pensar na soma desses desvios, mas, como para qualquer conjunto de dados,

$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, ou seja,

$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$ (verifique isto usando os conjuntos de dados acima), a opção seria considerarmos a soma dos quadrados dos desvios:

$SQD = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

O uso deste total, no entanto, pode causar dificuldades quando se comparam conjuntos de dados com números diferentes de observações. Deste modo, exprime-se esta medida como média, ou seja, a variância, que nada mais é do que dividir a *SQD* pelo número de observações da amostra (*n*).

2.2.1 Variância (σ^2 ou Var)

Considerando, então, a soma de quadrados dos desvios em relação à média, se estabelece uma medida de variabilidade para um conjunto de dados, chamada variância e definida como:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}, \quad \text{onde } X = x_i, \ i = 1, 2, ..., n$$

Será visto na aula "Estatística e distribuição amostral" que a variância de uma amostra é calculada, por motivos associados à inferência estatística, usando ***n-1*** em lugar de ***n*** nessa expressão, no entanto, para grandes amostras, pouca diferença fará o uso de *n* ou *n-1*. Portanto, a variância amostral é calculada pela fórmula:

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

Voltando ao Exemplo:

$$\sigma^2(A) = \frac{(25-31)^2 + (28-31)^2 + ... + (37-31)^2}{5} = \frac{90}{5} = 18{,}0$$

$$\sigma^2(B) = \frac{(17-31)^2 + (23-31)^2 + ... + (46-31)^2}{5} = \frac{550}{5} = 110{,}0$$

Então, podemos dizer que o grupo A é mais homogêneo que o B.

Fórmula da variância sem utilizar os desvios:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

$$\sigma^2 = \frac{1}{n}\left(\sum x_i^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2\right)$$

$$\sigma^2 = \frac{1}{n}\left(\sum x_i^2 - 2\bar{x}\frac{n}{n}\sum x_i + n\bar{x}^2\right)$$

$$\sigma^2 = \frac{1}{n}\left(\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right)$$

$$\sigma^2 = \frac{1}{n}\left(\sum x_i^2 - n\bar{x}^2\right)$$

$$\sigma^2 = \frac{1}{n}\left(\sum x_i^2 - n\left(\frac{\sum x_i}{n}\right)^2\right)$$

$$\sigma^2 = \frac{1}{n}\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)$$

Finalmente, a fórmula da variância sem a necessidade do cálculo dos desvios.

$$\sigma^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n}$$

Se n_i representa a frequência da observação x_i, i = 1, 2, ..., k, então podemos definir a variância como:

$$\sigma^2 = \frac{\left[\sum_{i=1}^{k} n_i (x_i - \bar{x})^2\right]}{n} = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2 \qquad \text{onde:} \quad n = \sum_{i=1}^{k} n_i \quad \text{e} \quad f_i = n_i / n$$

Desenvolvendo (3), obtêm-se:

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{k} n_i (x_i - \bar{x})^2\right]$$

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{k} n_i (x_i^{\,2} - 2x_i\bar{x} + \bar{x}^2)\right]$$

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{k} n_i x_i^{\,2} - 2\bar{x}\sum_{i=1}^{k} n_i x_i + \sum_{i=1}^{k} n_i \bar{x}^2\right]$$

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{k} n_i x_i^{\,2} - 2\bar{x}\frac{n}{n}\sum_{i=1}^{k} n_i x_i + n\bar{x}^2\right]$$

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{k} n_i x_i^{\,2} - 2\bar{x}n\sum_{i=1}^{k} f_i x_i + n\bar{x}^2\right], \text{onde}: \ \sum_{i=1}^{k} f_i x_i = \bar{x}$$

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{k} n_i x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right]$$

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{k} n_i x_i^2 - n\bar{x}^2\right]$$

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i x_i^2}{n} - \bar{x}^2$$

$$\sigma^2 = \sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2$$

Sendo a variância uma medida que expressa um desvio quadrático médio, pode causar alguns problemas de interpretação, uma vez que a unidade dos dados fica elevada ao quadrado. Para evitar isto, costuma-se usar o desvio padrão.

2.2.2 Desvio padrão (σ)

É definido como a raiz quadrada positiva da variância, ou seja

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \quad \text{ou} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n}}$$

Para amostras temos:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad \text{ou} \quad s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}}$$

Note que a unidade de medida do desvio padrão será a mesma dos dados originais. Temos, então, uma medida (básica) de variabilidade expressa na mesma unidade dos valores do conjunto de dados. Para o grupo A o desvio padrão é: $\sqrt{18,0} = 4,24$ e para o B: $\sqrt{110,0} = 10,49$.

O desvio padrão não é uma medida **resistente**. No caso do exemplo, onde foi mostrado que a mediana é uma medida resistente, utilizando-se as observações 5, 7, 8, 10 e 12, obtêm-se s = 2,41. Após a mudança de 12 para 120, obtêm-se 45,03, isto é, mais de 18 vezes a anterior; enquanto que a mediana não muda.

Exemplo 8. Calculemos a variância e o desvio padrão para a variável nota de C, P ou M (Tabela 2):

$$s = [\sum_{i=1}^{k} n_i(x_i - \bar{x})^2]/(n-1) = \frac{8(5-7,28)^2 + 32(7-7,28)^2 + 10(10-7,28)^2}{49} = 2,40$$

$$s = \sqrt{2,40} = 1,55$$

2.2.3 Medidas de dispersão para dados agrupados

O cálculo das medidas de dispersão, neste caso, é feito de modo análogo àquele usado para encontrar a média, ou seja, considerando-se que todas as observações no intervalo de classe, estão localizadas no ponto médio do intervalo. Para exemplificar, consideremos a Tabela 3, onde:

x_i (ponto médio)	146	172	198	224	250
$(x_i - \bar{x})$	-51,5	-25,5	0,5	26,5	52,5
n_i	6	11	17	10	6
$n_i (x_i - \bar{x})^2$	15901,1	7141,5	4,6	7033,1	16550,1

$i = 1, 2, \ldots, k$, $k = 5$ classes $\bar{x} = 197{,}48$ kg $n = 50$

$$s^2 = \sum_{i=1}^{k} n_i (x_i - \bar{x})^2 / (n-1) = (15901{,}1 + \ldots + 16550{,}1) / 50 = 951{,}6\ kg^2$$

$$s = \sqrt{951{,}6} = 30{,}8\ kg$$

Obs. Usando os dados da Tabela 1, $s^2 = 984{,}1\ kg^2$ e $s = 31{,}4\ kg$.

2.2.4 Coeficiente de variação *(CV)*

O desvio padrão, apesar de sua utilidade como medida de variabilidade, deve ser usado com cuidado, quando se compara variabilidades de diferentes conjuntos de dados. Por exemplo, um desvio padrão igual a 2 anos, seria considerado pequeno, se obtido em indivíduos com idade média igual a 55 anos, mas seria considerado grande se calculado em indivíduos com idade média igual a 3 anos. Além disso, o desvio padrão tem magnitude que é dependente da magnitude dos dados. Suínos ao abate, têm pesos que são, talvez, 50 vezes maiores do que de coelhos. Se os pesos dos suínos não forem mais variáveis que os dos coelhos, em relação às suas respectivas médias, o desvio padrão dos pesos dos suínos seria, mesmo assim, 50 vezes maior do que o dos coelhos (e a variância seria 50^2 = 2.500 vezes maior).

O coeficiente de variação, por sua vez, é uma medida de variação relativa, a qual expressa o desvio padrão como uma porcentagem da média ($\bar{x}$), ou seja, é o desvio padrão expresso em unidades de $\bar{x}$ (em %). Assim, o coeficiente de variação é definido como:

$$CV = 100 \frac{s}{\bar{x}}$$

com $\bar{x} \neq 0$,

o qual é interpretado como a variabilidade dos dados em relação à média. Como a razão $s/\bar{x}$, geralmente, é de pequeno valor, então, ela é multiplicada por 100 para expressá-la como uma porcentagem.

Voltando ao exemplo das idades, suponha dois grupos de indivíduos, sendo que em um deles os indivíduos têm idades 3, 1 e 5 anos e no outro, têm idades 55, 57 e 53 anos. No primeiro grupo, a média de idade é 3 anos e, no segundo grupo, a média é de 55 anos. Nos dois grupos a dispersão de idades é a mesma (s = 2), mas o desvio de dois anos é muito mais importante no primeiro grupo. Por quê?

Basta calcular o *CV* para cada um dos grupos. Para o primeiro grupo, o *CV* é 66,7%, ($2/3 \times 100$), enquanto que para o segundo grupo o CV é 3,6% ($2/55 \times 100$). Assim, desvios de 2 anos são muito mais importantes para o primeiro grupo que para o segundo, isto é, a dispersão dos dados em torno da média é muito grande no primeiro grupo. Como a média e o desvio padrão são expressos na mesma unidade de medida, o coeficiente de variação é adimensional (independe da magnitude ou da unidade de medida dos dados). Por exemplo, se os pesos aos 12 meses (P12), mostrados na Tabela 1 do capítulo anterior, estivessem sido medidos em gramas, em vez de quilogramas, o valor do *CV* dessa variável não se alteraria (veja cálculo no exemplo que se segue, com os dados em kg). Deste modo, o *CV* pode ser usado como um índice de variabilidade, sendo que sua grande utilidade é permitir a comparação das variabilidades de diferentes conjuntos de dados.

Exemplo 9. As variáveis Nota de C, P ou M (Tabela 2, Aula 1) e peso aos 12 meses (Tabela 3, Aula 1) deram os seguintes resultados:

Variável	$\bar{x}$	s
Nota	7,28	1,54
P12	197,48 kg	30,50 kg

Portanto, os coeficientes de variação dessas variáveis são, respectivamente,

$1{,}54 / 7{,}28 \times 100 = 21{,}2\%$ e $30{,}50 / 197{,}48 \times 100 = 15{,}4\%$,

os quais implicam que os desvios padrões das notas e dos pesos são 21,2% e 15,4% das respectivas médias. Assim, P12 se apresenta relativamente mais estável, embora o desvio padrão dos pesos seja 20 vezes maior do que o das notas.

Em resumo, se existirem dois conjuntos de observações distintos A e B, e se deseja saber qual deles é o mais homogêneo, ou seja, de menor variabilidade, basta fazer o seguinte: calculam-se as médias e os desvios padrões de A e B, e:

- se $\overline{X}_A = \overline{X}_B$, então o desvio padrão informará qual é o mais homogêneo
- se $\overline{X}_A \neq \overline{X}_B$, então o mais homogêneo será o que apresentar menor CV

OBS Valores muito altos de CV indicam pequena representatividade da média.

3 PROBABILIDADE

O termo **experimento** significa fazer ou observar alguma coisa sob certas condições, resultando em algum estado final de acontecimentos ou resultados. Na prática, os experimentos não são precisamente repetíveis, mesmo sob condições supostamente idênticas. Este é o caso quando há fatores afetando os resultados, mas não há conhecimento desses fatores ou como controlá-los e ainda quando há fatores supostamente sob controle, mas que na realidade não estão. Os resultados, então, não podem ser preditos a partir do conhecimento das "condições" (aquelas levadas em consideração), sob as quais o experimento é executado. Trata-se de um experimento envolvendo eventualidade ou, simplesmente, **experimento aleatório**.

Como o resultado do experimento não pode ser predito, é um de muitos resultados possíveis, um modelo que o represente deve incluir uma relação desses resultados. O conjunto de resultados possíveis é o **espaço amostral** do experimento. O segundo e principal componente de um modelo para um experimento aleatório é o conhecimento de **probabilidade**, que formaliza o conceito de que alguns conjuntos de resultados são mais ou menos frequentes do que outros.

3.1 Espaço amostral e Evento

Exemplo. Seja **A** um locus com dois alelos, **A** (dominante) e **a** (recessivo). Supondo os cruzamentos parentais Aa × Aa, os genótipos resultantes possíveis são:

M \ F	A	a
A	AA	Aa
a	aA	aa

Definição 1. O conjunto de todos os resultados possíveis associados com um experimento é chamado e**spaço amostral** (Ω ou U) do experimento.

Definição 2. Cada resultado possível é chamado de **ponto amostral** ou **evento elementar** ou **resultado elementar** (e_i).

$\Omega = \{e_1, e_2, ...\}$. No caso do exemplo acima, Ω = {AA, Aa, aA, aa}

Quando o espaço amostral contém um número finito, ou infinito, porém contável, de pontos, é chamado **espaço amostral discreto**. Se consiste de todos os números reais de determinado intervalo, é um **espaço amostral contínuo**.

Definição 3. Qualquer subconjunto, E, no espaço amostral Ω (ou em outras palavras, qualquer coleção de resultados elementares) é chamado **evento**.

Exemplo. E = descendente é dominante $(\mathbf{A}\_)$ = {AA, Aa, aA}

Nota: E = $\{e_1\}$ → evento simples

E = Ω → evento certo
E = ∅ → evento impossível

Para fins de facilitar a descrição, os princípios básicos de probabilidade serão mostrados aqui no contexto de espaços amostrais, tendo um número de eventos (ou resultados) elementares finito.

3.2 Probabilidade de um evento [*P(E)*]

Intuitivamente, pode ser definida como uma medida numérica com a qual se avalia "quão provável" é a ocorrência do evento, quando o experimento é executado. Para quantificar a expressão "quão provável" é natural tomar a fração de vezes que o evento ocorre em repetidas tentativas do experimento. Assim, o conceito intuitivo de uma medida numérica para a probabilidade de um evento é em termos da ***proporção de vezes que o evento é esperado ocorrer***, quando o experimento é repetido sob idênticas condições. O processo apropriado para se determinar probabilidades para eventos depende da natureza do experimento e do espaço amostral associado. Há dois tipos de situações:

3.2.1 **Resultados elementares igualmente prováveis**

Em alguns casos, a proporção de vezes que cada resultado elementar é esperado ocorrer pode ser determinado sem executar o experimento. Assim, se um espaço amostral Ω consiste de *k* resultados elementares $\{e_1, e_2 ..., e_k\}$ que são igualmente prováveis de ocorrerem, a probabilidade de cada e_i é **1/k**. Se um evento E consiste de **m** desses **k** elementos, então:

$$P(E) = \frac{m}{k} = \frac{\text{Número de elementos em E}}{\text{Número de elementos em }\Omega}$$

Exemplo. $P(\text{descendente é dominante}) = P(A\_) = \frac{3}{4}$

Nesta condição, não é necessário explicitar completamente Ω e E para se calcular *P(E)*, basta calcular *m* e *k*. Para tanto, são usados os métodos clássicos de contagem da análise combinatória. Um princípio fundamental de contagem diz que, se uma tarefa pode ser executada em duas etapas, a primeira podendo ser realizada de ***p*** maneiras e a segunda de ***q*** maneiras, então, a tarefa completa pode ser executada de ***p×q*** maneiras.

Exemplo. Suponha que em um lote com 20 animais existem 5 doentes. Escolhem-se 4 animais do lote ao acaso, isto é, uma amostra de 4 elementos, de modo que a ordem dos elementos seja irrelevante. Considerando o evento *E*: 2 doentes na amostra, calcular *P(E)*.

$k = \binom{20}{4}$ é o número de amostras com 4 elementos que pode-se extrair do lote (número de pontos do espaço amostral).

$M = \binom{5}{2}\binom{15}{2}$ é o número de maneiras que pode-se escolher 2 doentes e 2 não doentes, simultaneamente, na amostra de 4 elementos

$$P(E) = \quad P(E) = \frac{{}_{5}C_{2}\;{}_{15}C_{2}}{{}_{20}C_{4}} = \frac{\binom{5}{2}\binom{15}{2}}{\binom{20}{4}} = \frac{\frac{5!}{3!2!}\times\frac{15!}{13!2!}}{\frac{20!}{16!4!}} = \frac{10\times 105}{4845} \cong 0{,}22$$

Sendo *E:* 4 doentes na amostra

$$P(E) = \frac{\binom{5}{4}\binom{15}{0}}{\binom{20}{4}} = \frac{5}{4845} \cong 0{,}001$$

3.2.2 Probabilidade e frequência relativa

Em outras situações, é necessário repetir o experimento um grande número de vezes para se obter informações a respeito da frequência de ocorrência dos diferentes resultados. Por exemplo, a razão fenotípica Dominantes:Recessivos = 3:1 foi primeiro deduzida por Mendel, com base nos resultados do seu experimento clássico de cruzamentos para cor de sementes de ervilhas:

P AA (amarelas) x aa (verdes)

F_1 Aa (amarelas)

$F_1 \times F_1 \rightarrow F_2$ (amarelas e verdes)

Em F_2, ele observou a razão:

$$\frac{\textit{Número de plantas com sementes amarelas}}{\textit{Número de plantas no experimento}}$$

Tal razão é chamada frequência relativa. Repetindo o experimento várias vezes, Mendel observou que a mesma aproximou-se de um limite igual a ¾.

Em geral, quando um experimento é repetido ***n*** vezes, define-se como frequência relativa de um evento *E* em ***n*** ensaios a razão:

$$f_n(E) = \frac{\textit{Número de vezes que } E \textit{ ocorre em } n \textit{ ensaios}}{n}$$

A razão $f_n(E)$ flutua quando o número $\boldsymbol{n}$ de repetições do experimento muda. Entretanto, desde que as condições experimentais não mudem, a $f_n(E)$, quando $\boldsymbol{n}$ aumenta ($n \to \infty$), tende a se estabilizar em um valor numérico único, o qual é chamado de ***probabilidade do evento E***. Este comportamento é ilustrado na Figura 1.

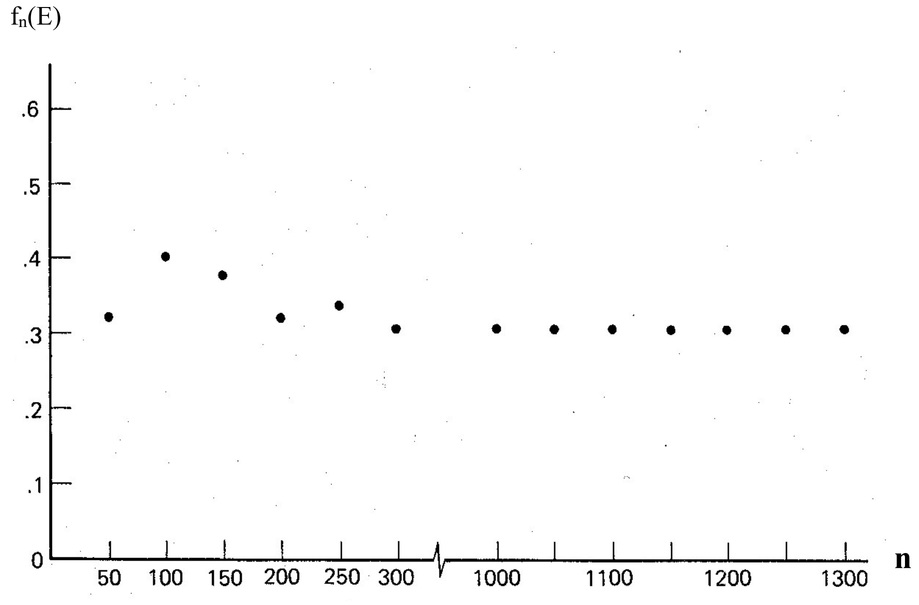


Figura 1. Estabilização da frequência relativa

3.2.3 Algumas propriedades

Como toda frequência relativa é um número entre 0 e 1,

$$0 < P(E) < 1$$

para qualquer evento E. Considerando o espaço amostral (Ω) e o conjunto vazio ($\varnothing$) como eventos, temos $P(\Omega) = 1$ e $P(\varnothing) = 0$.

Exemplo. Suponha que o quadro seguinte represente a distribuição dos animais de um dado rebanho.

	Sexo		
Raça	Macho (M)	Fêmea (F)	Total
Nelore (N)	70	40	110
Guzerá (G)	15	15	30
Canchim (C)	10	20	30
Indubrasil (I)	20	10	30
Total	115	85	200

Indicando por G o evento que ocorre, quando se escolhendo ao acaso um animal, ele for da raça Guzerá (N, C, I, M e F têm significados análogos), então:
$P(G) = 30/200$ e $P(M) = 115/200$

Dados os eventos G e M, podem-se considerar dois novos eventos:

(1) $G \cup M$, chamado **reunião ou união** de G e M, que ocorre quando pelo menos um dos eventos ocorre; e

(2) $G \cap M$, chamado **intersecção** de G e M, que ocorre quando G e M ocorrem simultaneamente.

No exemplo:

$P(G \cap M) = 15/200$ e

$$P(G \cup M) = P(G) + P(M) - P(G \cap M) = \frac{30}{200} + \frac{115}{200} - \frac{15}{200} = \frac{130}{200}$$

Considerando-se, no entanto, os eventos G e I,

$$P(G \cup I) = P(G) + P(I) = \frac{30}{200} + \frac{30}{200} = \frac{60}{200}$$

Neste caso, os eventos G e I são **mutuamente exclusivos** ou **disjuntos**, isto é, a ocorrência de G exclui a ocorrência de I e vice-versa. Assim sendo,

$G \cap I = \varnothing$ e $P(G \cap I) = 0$

Portanto, se A e B são dois eventos quaisquer, tem-se a chamada **regra da adição de probabilidades**:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, que se reduz a

$P(A \cup B) = P(A) + P(B)$, se A e B são **disjuntos**

Para três eventos, A_1, A_2 e A_3, têm-se:

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) -$$

$$- P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$$

Esta relação pode ser estendida para um número finito qualquer de eventos.

Evento complementar

O evento consistindo dos pontos amostrais em Ω que não pertencem a um evento E é chamado **complemento** de E, e é indicado por $\overline{E}$ ou E^C.

$P(\overline{E}) = 1 - P(E)$ $\Omega = E \cup \overline{E}$

Como $E \cap \overline{E} = \varnothing$, $P(\Omega) = P(E) + P(\overline{E}) = 1$, logo

$$P(\overline{E}) = 1 - P(E)$$

Esta relação pode ser usada para calcular $P(\overline{E})$, quando E é simples e $P(E)$ é facilmente calculada.

Exemplo. Sejam os eventos **G** e **A** = N ∪ C ∪ I, onde G ∪ A = Ω e G ∩ A = ∅. Portanto, G e A são **complementares**.

Vimos que $P(G)$ = 30/200, enquanto que

$P(A) = 110/200 + 30/200 + 30/200 = 170/200$. Isto é,

$P(G) + P(A)$ = 1, então $P(\overline{G}) = 1 - P(G) = P(A)$

3.3 Probabilidade condicional e independência de eventos

Considerando (dado) agora que o animal escolhido ao acaso é da raça Canchim (C), a probabilidade de que seja fêmea (F) é 20/30 = 2/3. Escreve-se:

$P(Fêmea \mid Canchim) = 20/30 = 2/3$

Para dois eventos quaisquer, A e B, a probabilidade de A quando se sabe que B ocorreu, é chamada **probabilidade condicional de A dado B**, P(A | B), e é calculada por:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad (1)$$

desde que $P(B) > 0$

Para o exemplo mencionado,

$P(C)$ = 30/200 e $P(F \cap C)$ = 20/200, então

$$P(F \mid C) = \frac{P(F \cap C)}{P(C)} = \frac{20/200}{30/200} = 2/3\text{, como obtido.}$$

As propriedades acima e a probabilidade condicional podem ser apresentadas nas formas de diagramas, como mostrado na Figura1.

Da relação (1), obtêm-se a chamada **regra do produto de probabilidades**:

$P(A \cap B) = P(B) \times P(A \mid B) = P(A) \times P(B \mid A)$

Se $P(A\,|\,B) = P(A)$, $P(B\,|\,A) = P(B)$, isto é, se a probabilidade de ocorrência de A (ou de B) não é afetada pela ocorrência, ou não de B (ou de A), os dois eventos se dizem **independentes**. Neste caso,

$$P(A \cap B) = P(A) \times P(B) \qquad (2)$$

Reciprocamente, se (2) verifica-se, A e B são **independentes**.

Vejamos agora o conceito de independência para três eventos. Se A_1, A_2 e A_3 são **independentes**, então eles devem ser independentes dois a dois

$$P(A_j \cap A_k) = P(A_j) \times P(A_k) \qquad j \neq k \quad \text{onde: } j, k = 1, 2, 3 \qquad (3)$$

e também $P(A_1 \cap A_2 \cap A_3) = P(A_1) \times P(A_2) \times P(A_3)$ (4)

Nenhuma das expressões (3) ou (4) é por si só suficiente. É fácil generalizar para mais de três eventos.

Exemplo. Um grupo de pessoas foi classificado quanto a peso e pressão arterial, apresentando as proporções do quadro a seguir:

	Peso			
Pressão	Excesso (B)	Normal	Deficiente	Total
Elevada (A)	0,10	0,08	0,02	0,20
Normal	0,15	0,45	0,20	0,80
Total	0,25	0,53	0,22	1,00

Verifique se os eventos A e B são independentes ou não.

$$P(A) = 0{,}20 \qquad\qquad P(A\,|\,B) = \frac{P(A \cap B)}{P(B)} = \frac{0{,}10}{0{,}25} = 0{,}40$$

Portanto, $P(A) \neq P(A\,|\,B)$, isto é, os eventos A e B **não são independentes**. Alternativamente, $P(A \cap B) \neq P(A) \times P(B)$

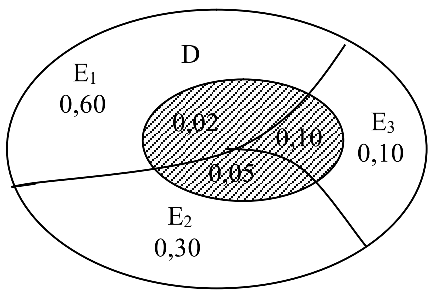
3.4 Teorema de Bayes

Para ilustrá-lo, consideremos o seguinte exemplo: em um rebanho, tem-se E_i = lotes de animais e D = animais doentes, em que:

$P(D|E_1) = 0{,}02$

$P(D|E_2) = 0{,}05$

$P(D|E_3) = 0{,}10$



Toma-se um lote ao acaso e dele retira-se um animal. **É doente**. Qual a probabilidade do lote escolhido ser E_1, ou seja, $P(E_1 \mid D)$?

Solução: Da definição de probabilidade condicional, temos

$$P(E_1 \mid D) = \frac{P(E_1 \cap D)}{P(D)}$$

O numerador dessa expressão pode ser reescrito pela regra do produto, condicionado à E_1, isto é, $P(E_1 \cap D) = P(E_1) \times P(D \mid E_1)$, tal que

$$P(E_1 \mid D) = \frac{P(E_1 \cap D)}{P(D)} = \frac{P(E_1) \times P(D / E_1)}{P(D)} \qquad (1)$$

Assim, precisamos encontrar o valor de $P(D)$, já que o numerador é conhecido. Como E_1, E_2 e E_3 são eventos mutuamente exclusivos, e reunidos formam o espaço amostral completo (Ω), podemos decompor o evento D na reunião de três outros, também mutuamente exclusivos, como segue:

$D = (E_1 \cap D) \cup (E_2 \cap D) \cup (E_3 \cap D)$, e então

$P(D) = P(E_1 \cap D) + P(E_2 \cap D) + P(E_3 \cap D)$

Substituindo $P(D)$ em (1), obtemos

$$P(E_1 \mid D) = \frac{P(E_1) \times P(D \mid E_1)}{P(E_1 \cap D) + P(E_2 \cap D) + P(E_3 \cap D)}$$

Reescrevendo o denominador dessa expressão pela regra do produto, condicionado à E_i, para i = 1, 2 e 3, temos

$$P(E_1 \mid D) = \frac{P(E_1) \times P(D / E_1)}{P(E_1) \times P(D \mid E_1) + P(E_2) \times P(D \mid E_2) + P(E_3) \times P(D \mid E_3)} \qquad (2)$$

do que segue que

$$P(E_1 \mid D) = \frac{0{,}6 \times 0{,}02}{0{,}6 \times 0{,}02 + 0{,}3 \times 0{,}05 + 0{,}1 \times 0{,}10} = 0{,}32$$

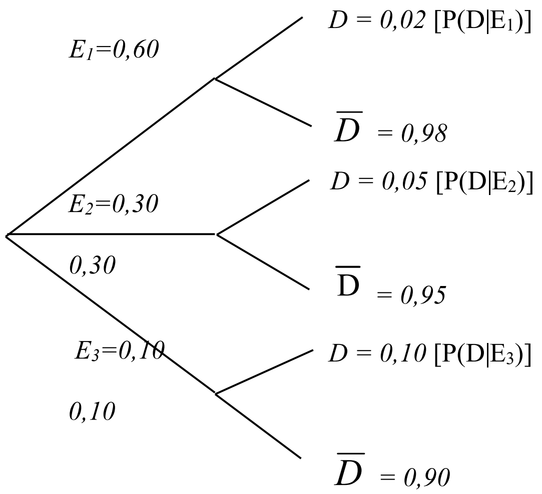
Esse resultado (2) pode ser generalizado do seguinte modo: seja E_1, E_2, …, E_k uma sequência de eventos mutuamente exclusivos, com probabilidades $P(E_1)$, $P(E_2)$, ..., $P(E_k)$, respectivamente; e D um evento que ocorre, com $P(D) > 0$, quando e somente quando um dos eventos E_1, E_2, …, E_k ocorre. Os eventos E_1, E_2, …, E_k determinam as diferentes condições ou causas sobre os quais D pode ocorrer. As probabilidades $P(E_1)$, $P(E_2)$, ..., $P(E_k)$ são chamadas probabilidades *a priori* da ocorrência desses eventos, sem levar em conta o evento D.

Seja $P(D|E_i)$, $i = 1, 2, \ldots, k$, a probabilidade condicional de ocorrência do evento D, dado que o evento E_i tenha ocorrido. Devemos assumir que as probabilidades $P(E_i)$ e $P(D|E_i)$, $i = 1, 2, \ldots, k$, são conhecidas.

Desejamos encontrar a probabilidade do evento E_i, supondo a ocorrência do evento D, isto é, $P(E_i|D)$, chamada probabilidade *a posteriori* de Ei, calculada depois que D tenha sido observado. A fórmula com a qual $P(Ei|D)$ pode ser calculada:

$$P(E_i|D) = \frac{P(E_i) \times P(D|E_i)}{\sum_{j=1}^{k} P(E_j) \times P(D|E_j)} \quad \textit{para todo } i = 1, 2, \ldots, k \qquad (3)$$

é conhecida como **Teorema de Bayes**, que expressa uma probabilidade condicional em termos de outras probabilidades condicionais e marginais.
Essas probabilidades podem ser teoricamente deduzidas a partir de um modelo representado pelo espaço amostral em que esses eventos são definidos. A visualização do problema é facilitada pela utilização do **Diagrama em Árvore**, ilustrado a seguir usando os dados do exemplo:



De modo que, pelo Teorema de Bayes, temos

$$P(E_1|D) = \frac{0,6 \times 0,02}{0,6 \times 0,02 + 0,3 \times 0,05 + 0,1 \times 0,10} = \frac{0,12}{0,037} = 0,3243 = 32,43\%$$

$$P(E_2|D) = \frac{0,05 \times 0,30}{0,037} = 0,4054(40,54\%)$$

$$P(E_3|D) = \frac{0,10 \times 0,10}{0,037} = 1 - (0,3243 + 0,4054) = 0,2703(27,03\%)$$

4 VARIÁVEIS ALEATÓRIAS

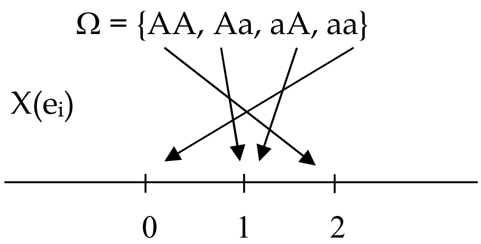
Uma variável cujos valores referem-se a eventos aleatórios é chamada **variável aleatória**; seus valores dependem dos resultados de um experimento. Pode ser **discreta** ou **contínua.**

Variável aleatória discreta

4.1.1 **Definição**

Muitos experimentos produzem resultados **não numéricos**. Antes de analisá-los é conveniente transformar seus resultados em números. Para isso devemos associar a cada resultado elementar (e_i) do espaço amostral (Ω) um número real, o que é feito por meio de uma regra ou função denominada **variável aleatória**.

Exemplo 1. Considerando o cruzamento Aa × Aa, este conceito é ilustrado com um espaço amostral com 4 resultados elementares, ou seja:



onde X denota o número de genes A no genótipo. Assim definida, X é uma variável aleatória.

Note que para ser discreta, a variável aleatória (v.a.) deve assumir valores em um conjunto finito ou infinito, porém contável.
O passo fundamental para entendermos uma v.a. é associar a cada valor de X sua probabilidade, obtendo o que se chama uma **distribuição de probabilidade**.

4.1.2 **Distribuição de probabilidade**

Definição. É uma relação dos distintos valores x_i de X junto com as suas respectivas probabilidades $p(x_i)$, com $\sum_i p(x_i) = 1$.

Exemplo 2. Considerando os descendentes de Aa × Aa, a distribuição do número de genes A nos genótipos (X) é idêntica à distribuição de genótipos, ou seja

Genótipos	AA	Aa	aa	Total
$X = x_i$	2	1	0	
$P(X = x_i) = p(x_i)$	1/4	1/2	1/4	1,0

em que: $p(x_i)$ é chamada **função de probabilidade**, que a cada valor de x_i associa sua probabilidade de ocorrência.

A distribuição de probabilidade mostra-nos como a probabilidade total (1,0) é distribuída de acordo com os diferentes valores da variável aleatória.
Frequentemente, uma fórmula matemática pode ser usada para representar, em lugar de uma tabela, uma distribuição de probabilidade.

4.1.3 Representação gráfica de uma distribuição de probabilidade

Gráfico de barras

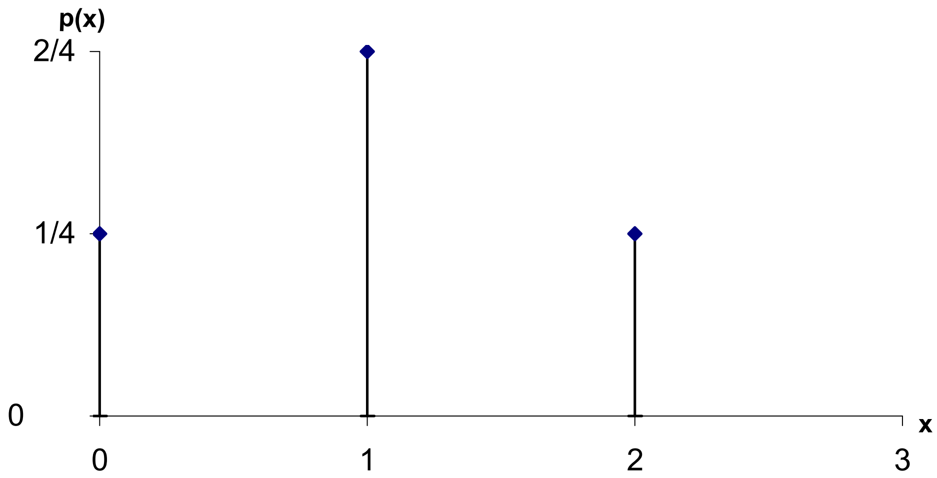
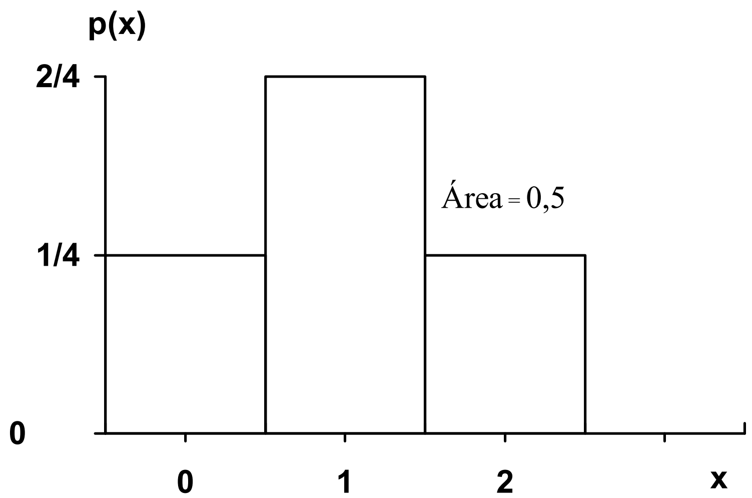


Gráfico de barras para a distribuição dada no Exemplo 2

(b) Histograma

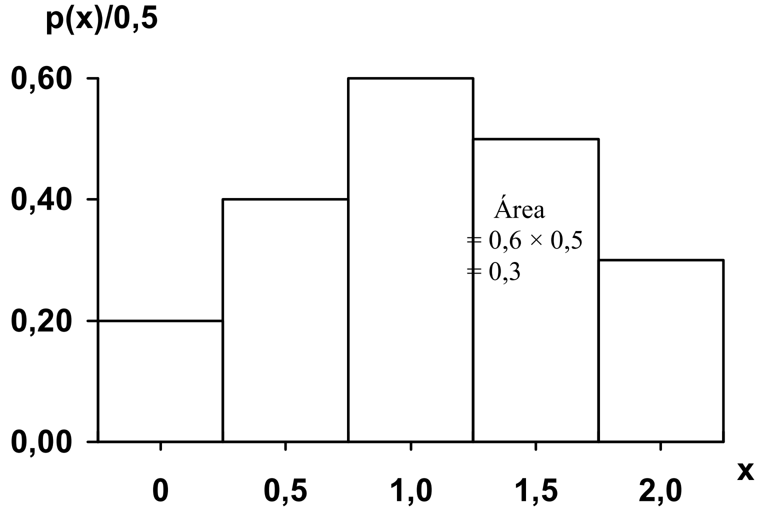


Histograma para a distribuição dada no Exemplo 2

Quando o espaçamento entre os valores de X difere de 1,0, tal como na seguinte distribuição de probabilidade.

X	0	0,5	1,0	1,5	2,0
$p(x)$	0,1	0,2	0,3	0,25	0,15

histograma é traçado como:



Ou seja, as alturas dos retângulos são determinadas dividindo-se as probabilidades $p(x)$ pelas bases dos mesmos.
O histograma é recomendado para distribuições com valores de X igualmente espaçados. Caso contrário, o gráfico de barras deve ser usado.

4.2 Esperança matemática

Exemplo 3. Seja uma **população finita** de n indivíduos

Genótipos	AA	Aa	aa	Total
Número	n_1	n_2	n_3	n
$X = x_i$	2	1	0	

Denotando X o número de genes A no genótipo, o **número médio** de genes A ($\bar{x}$) é:

$$\bar{x} = \frac{1}{n}(2\times n_1 + 1\times n_2 + 0\times n_3) = 2\frac{n_1}{n} + 1\frac{n_2}{n} + 0\frac{n_3}{n}$$

Esta é a média para uma população finita de tamanho n. Considerando um modelo de ***população infinita***, as frequências relativas n_i/n (i = 1, 2, 3) podem se aproximar de limites que são probabilidades $P(X = x_i) = p(x_i)$, onde:
x_i = 2, 1, 0, e $\bar{x}$ se aproximará de um limite que é chamado ***Esperança de X*** (***<u>isto é, o número esperado de genes A em uma população infinita</u>***). O resultado pode ser generalizado na seguinte definição:

Definição. A média de uma v.a. ***X*** ou de sua distribuição de probabilidade, também chamada ***valor esperado*** ou ***esperança matemática*** ou simplesmente ***esperança de X, E(X)***, é definida como:

$$E(X) = \sum_{i=1}^{k} x_i \times p(x_i)$$

E(X) é usada como medida do centro da distribuição de probabilidade. Por isso, é também chamada média populacional e simbolizada por μ. Na verdade, *E(X)* é uma média hipotética que pode nunca ser observada, mas é "***esperada***" em uma população.

Exemplo 4. Usando a distribuição de probabilidade dada no Exemplo 2:

$$E(X) = 0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 2 \times \frac{1}{4} = 1$$

O número esperado de genes A nos descendentes de Aa × Aa é igual a 1.

4.2.1 Propriedades da esperança

Se ***a*** e ***b*** são constantes e *X* uma v.a., então:

i. $E(a) = a$
ii. $E(bX) = bE(X)$
iii. $E(X + a) = E(X) + a$
iv. $E(a + bX) = a + bE(X)$
v. $E(a + bX + cX^2) = a + bE(X) + cE(x^2)$

4.3 Variância

Definição. A **variância** de uma v.a. *X* ou a medida de dispersão de sua distribuição de probabilidade, representada por σ^2_X, é definida por

$$\sigma^2_X = Var(X) = E[(X - \mu)^2]$$

A variância pode ser calculada de dois modos:

(a) $E[(X-\mu)^2] = \sum_i (x_i - \mu)^2 p(x_i)$

(b) $E[(X-\mu)^2] = E(X^2) - \mu^2 = \sum x_i^2 p(x_i) - [E(X)]^2$

O ***desvio padrão*** (σ) é a raiz quadrada positiva da variância.

Exemplo 5. Seja a distribuição de probabilidade do Exemplo 2, então

$$\sigma^2_X = E[(X-\mu)^2] = (2-1)^2 \frac{1}{4} + (1-1)^2 \frac{2}{4} + (0-1)^2 \frac{1}{4} = \frac{1}{2} \text{ ou}$$

$$E[(X-\mu)^2] = (2^2 \frac{1}{4} + 1^2 \frac{2}{4} + 0^2 \frac{1}{4}) - 1^2 = \frac{1}{2}$$

4.3.1 Propriedades da variância

Para ***a*** e ***b*** denotando constantes e *X* uma v.a.,

i. *Var(X)* não pode ser negativa
ii. $Var(X + a) = Var(X)$
iii $Var(b.X) = b^2. Var(X)$

iv. $Var(a + b.X) = b^2. Var(X)$

Exemplo 6. Um revendedor de produtos veterinários recebe de vários laboratórios certo tipo de antibiótico, que tem custo diferenciado. Levando-se em conta a proporção fornecida e o preço apresentado por cada laboratório, pode-se considerar que o custo de uma dose de antibiótico em reais, escolhida ao acaso, é uma variável aleatória *C*. Admitindo a seguinte distribuição de probabilidade para *C*:

c_i	1,00	1,10	1,20	1,30	1,40
$p(c_i)$	0,2	0,3	0,2	0,2	0,1

Determinar a média e a variância da variável aleatória C:

$E(C) = 1{,}0\times0{,}2 + 1{,}1\times0{,}3 + 1{,}2\times0{,}2 + 1{,}3\times0{,}2 + 1{,}4\times0{,}1= 1{,}17$

$Var(C) = [(1{,}0^2\times0{,}2 + 1{,}1^2\times0{,}3 + 1{,}2^2\times0{,}2 + 1{,}3^2\times0{,}2 + 1{,}4^2\times0{,}1) - 1{,}17^2] = 0{,}016$

(b) Supondo que o revendedor venda cada um desses antibióticos acrescentando 50% sobre o custo, além de um adicional de R$ 0,10 pelo frete, calcular a média e a variância da nova variável aleatória preço de revenda *R*.

$r_i = 1{,}5c_i + 0{,}10$. Assim, usando as propriedades da média e da variância:

$E(R) = 1{,}5\times E(C) + E(0{,}10) = 1{,}5\times1{,}17 + 0{,}10 = 1{,}855$

$Var(R) = 1{,}5^2\times Var(C) = 1{,}5^2\times0{,}016 = 0{,}036$

4.4 Distribuições teóricas de probabilidades de variáveis aleatórias discretas

Nas diversas áreas de pesquisa é comum o aparecimento de variáveis aleatórias discretas, como resultados de experimentos aleatórios. Assim, para um dado experimento, deve-se verificar se ele satisfaz as condições dos modelos probabilísticos conhecidos, pois isso facilitaria muito sua análise. Por modelo probabilístico para uma variável aleatória X, ***entende-se como uma forma específica de função de distribuição de probabilidade que reflita o comportamento de X***. Aqui, serão estudados alguns desses modelos, procurando enfatizar as condições em que aparecem, suas funções de probabilidades, parâmetros, e como calcular probabilidades.

4.4.1 Distribuição de Bernoulli

Consideremos uma única tentativa de um experimento aleatório, onde há somente dois resultados possíveis, designados por: Sucesso **(*S*)** e Fracasso **(*F*)**. O uso destes termos é sugerido apenas por conveniência e não têm a mesma conotação de sucesso e fracasso na vida real. Habitualmente, o resultado de interesse principal é rotulado como sucesso, mesmo que se trate de um evento indesejável. Por exemplo:

testa-se um antibiótico em um indivíduo, a reação ou é positiva (*S*) ou é negativa (*F*);
observa-se um nascimento, o recém-nascido ou é macho (F) ou é fêmea (*S*);

(c) um animal é escolhido, ao acaso, de um lote contendo 50 animais, o animal é doente (S) ou não (F).

Em todos estes casos, estaremos interessados na ocorrência de um sucesso ou fracasso. Assim, para cada experimento, podemos definir uma variável aleatória X: o número de sucessos, que assume apenas dois valores, o valor 1 se ocorre sucesso (S) e o valor 0 (zero) se ocorre fracasso (F), sendo $P(S) = p$,
$0 < p < 1$. Ou seja:

$X = \begin{cases} 0\,(F) \\ 1\,(S) \end{cases}$ com $P(X = 1) = p$ e $P(X = 0) = 1 - p = q$

Nestas condições, a variável aleatória X com a função de probabilidade:

X	0 1	Total
$p(x)$	q p	1,0

ou $P(X = x) = \boldsymbol{p^x} \times \boldsymbol{q^{1-x}}$

é chamada ***variável aleatória de Bernoulli.***
Experimentos que resultam numa variável aleatória de Bernoulli são chamados ***<u>ensaios de Bernoulli</u>.***

Esperança e variância

$$E(X) = \sum_{i=1}^{k} x_i\, p(x_i) = 0 \times q + 1 \times p = p$$

$$Var(X) = E(X^2) - [E(X)]^2 = (0^2 \times q + 1^2 \times p) - p^2 = p - p^2 = p\,(1 - p) = p \times q$$

4.4.2 Distribuição Binomial

Quando um número fixo ***n*** de ensaios de Bernoulli são repetidos, supondo que as repetições sejam **independentes** (isto é, o resultado de um ensaio não tem influência no resultado de qualquer outro), com $P(S) = p$ em cada ensaio, pode-se considerar a variável aleatória X, que representa a contagem do número de sucessos em n ensaios. Os possíveis valores de X são os inteiros 0, 1, 2,..., n. A distribuição de probabilidade de X é chamada **distribuição binomial** com n ensaios e probabilidade de sucesso p.

Para deduzir uma fórmula para $P(X = x)$, onde $x = 0, 1, 2, ..., n$, ou seja x pode ser qualquer número inteiro entre 0 e n, consideremos $n = 4$ ensaios, cada um dos quais podendo resultar em S ou F. Há 2 × 2 × 2 × 2 = 16 resultados possíveis, os quais estão relacionados nas colunas abaixo, de acordo com o número de sucessos (S):

FFFF	SFFF	SSFF	SSSF	SSSS
	FSFF	SFSF	SSFS	
	FFSF	SFFS	SFSS	
	FFFS	FSSF	FSSS	
		FSFS		
		FFSS		

Valor de X (número de S)	0	1	2	3	4
Prob. de cada sequência	q^4	pq^3	p^2q^2	p^3q	p^4
Número de sequências	$1=\binom{4}{0}$	$4=\binom{4}{1}$	$6=\binom{4}{2}$	$4=\binom{4}{3}$	$1=\binom{4}{4}$

Como os ensaios são independentes e em cada ensaio $P(S) = p$ e $P(F) = q$, a probabilidade de cada sequência, por exemplo na terceira coluna, que tem 2 S's e 2 F's é $P(SSFF) = P(S) \times P(S) \times P(F) \times P(F) = p^2q^2$. Da mesma maneira, a probabilidade de cada sequência individual nesta coluna é p^2q^2. Há seis sequências, assim obtêm-se $P(X = 2) = 6\, p^2q^2$. O fator 6 é o número de
sequências com 2 S's e 2 F's. Mesmo sem fazer uma listagem completa das sequências, pode-se obter esta contagem, notando que os dois lugares onde S ocorre, podem ser selecionados de um total de 4 lugares em $\binom{4}{2} = 6$ maneiras, cada um dos remanescentes 2 lugares sendo sempre preenchidos com um F. Assim procedendo em relação às demais colunas, a distribuição binomial com $n = 4$ ensaios, pode ser disposta na forma da tabela apresentada a seguir:

Distribuição binomial com $\boldsymbol{n} = 4$ ensaios:

X	0	1	2	3	4
$P(X = x)$	$\binom{4}{0}p^0q^4$	$\binom{4}{1}p^1q^3$	$\binom{4}{2}p^2q^2$	$\binom{4}{3}p^3q^1$	$\binom{4}{4}p^4q^0$

Estendendo o raciocínio para o caso geral de $\boldsymbol{n}$ ensaios de Bernoulli, observa-se que há $\binom{n}{x}$ sequências que tem x sucessos e $(n - x)$ fracassos e que a probabilidade de cada sequência é $p^x.q^{n-x}$. Portanto,

$$P(X = x) = \binom{n}{x} p^x.q^{n-x} \quad \text{para x = 0, 1, 2,..., n}$$

Denota-se esta probabilidade por $\boldsymbol{b(x;\, n,\, p)}$, e quando X tem distribuição binomial com os parâmetros n e p escreve-se $\boldsymbol{X : b(n,\, p)}$.

O termo distribuição binomial é originado do "***teorema da expansão binomial***":

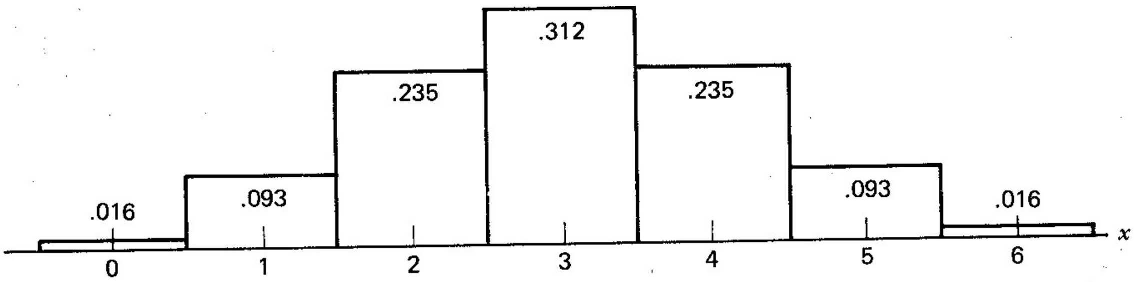
$$(a+b)^n = a^n + \binom{n}{1} ba^{n-1} + \binom{n}{2} b^2a^{n-2} + \ldots + \binom{n}{x} b^x a^{n-x} + \ldots + b^n$$

Considerando, em *particular*, $a = q$ e $b = p$, esta fórmula produz:

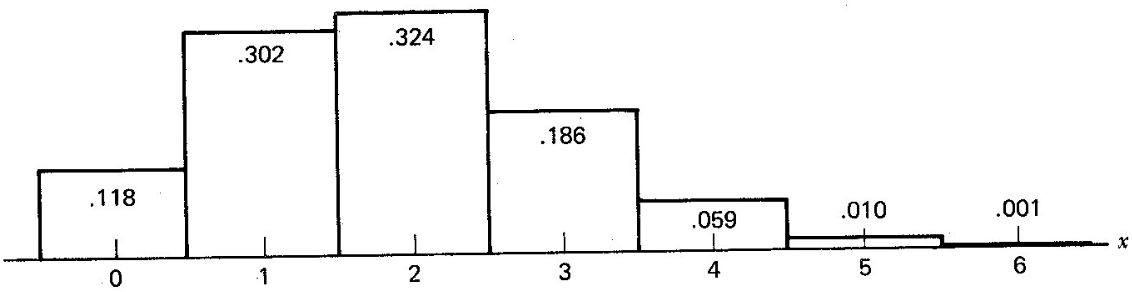
$$(q+p)^n = q^n + \binom{n}{1} pq^{n-1} + \binom{n}{2} p^2q^{n-2} + \ldots + \binom{n}{x} p^x q^{n-x} + \ldots + p^n$$

Os termos sucessivos do lado direito desta fórmula são as probabilidades binomiais. Como $p + q = 1$, $\sum_{x=0}^{n} b(x; n, p) = 1$, como seria para qualquer distribuição de probabilidades.

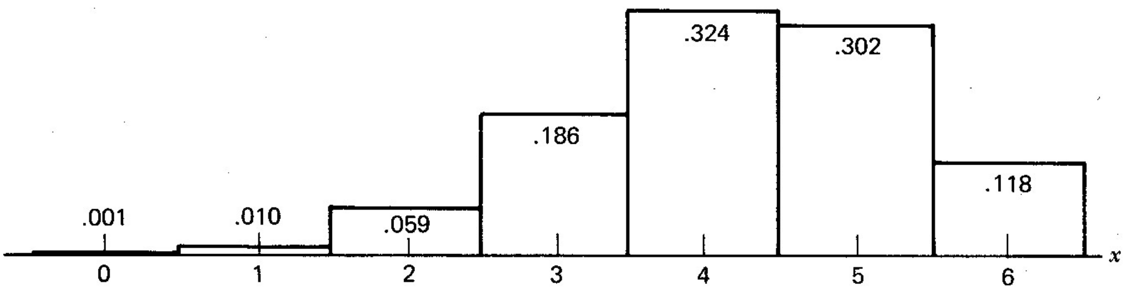
Ilustração da maneira pela qual os valores de $\boldsymbol{p}$ influenciam a forma da distribuição binomial:



(a) $n = 6$, $p = 0{,}5$ ($q = 0{,}5$)



(b) $n = 6$, $p = 0{,}3$ ($q = 0{,}7$)



(c) n = 6, $p = 0{,}7$ ($q = 0{,}3$)

Quando $p = 0{,}5$ (Figura a), a distribuição binomial é simétrica; se o valor de p em um histograma tem o mesmo valor de q em outro (Figuras b e c), as probabilidades são exatamente as mesmas, mas dispostas de forma invertida. Isto ilustra a propriedade geral da distribuição binomial: quando p e q são alternados, a distribuição de probabilidades é invertida. Então, pode-se estabelecer a relação geral $b\,(x; n, p) = b\,(n - x; n, 1 - p)$.

Uso da tabela binomial

A Tabela 1 apresenta os valores de $b(x; n, p)$ para **n** = 1 a 20 e **p** = 0,05; 0,10; 0,15; ...; 0,50. Quando p> 0,50, usa-se:

$\boldsymbol{b(x; n, p) = b(n - x; n, 1 - p)}$

Exemplificando, $b(2; 6, 0{,}7) = b(4; 6, 0{,}3) = 0{,}0595$

		p									
n	x	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
1	0	.9500	.9000	.8500	.8000	.7500	.7000	.6500	.6000	.5500	.5000
	1	.0500	.1000	.1500	.2000	.2500	.3000	.3500	.4000	.4500	.5000
2	0	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0312
	1	.2036	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1562
	2	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0000	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0000	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0938
	6	.0000	.0000	.0000	.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0000	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6	.0000	.0000	.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0037	.0078
8	0	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.2793	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0312
	2	.0515	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0054	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188
	4	.0004	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734

Tabela 1. Probabilidades binomiais

Esperança e Variância

A média e a variância de uma distribuição binomial são dadas por:

$E(X) = n.p$ e $Var(X) = n.p.q$

Para justificar essas fórmulas, consideremos que uma variável aleatória X que representa o número de sucessos em n ensaios de **Bernoulli** pode ser denotada por: $X = X_1 + X_2 + ...+ X_n$, onde X_i é o número de sucessos no i-ésimo ensaio (X_i = 0 ou 1). Como os ensaios são independentes, X_1, X_2, ..., Xn são variáveis aleatórias independentes, cada uma tendo distribuição de Bernoulli, em que $E(X_i) = p$ e $Var(X_i) = pq$.

Usando as propriedades de esperança e variância da soma de variáveis aleatórias, obtém-se:

$E(X) = E(X_1) + E(X_2) + ... + E(X_n) = p + p +... + p = n.p$

$Var(X) = Var(X_1) + Var(X_2) + ... + Var(X_n) = pq + pq +... + pq = n.p.q$

Exemplo 1. Ocorrendo 3 nascimentos a partir do acasalamento Aa × aa, qual a probabilidade de se obter 3 descendentes Aa?

P(Desc.Aa | Acas.Aa × aa) = p = 1/2

$$P(X = x) = \binom{n}{x} p^x \times q^{n-x} \Rightarrow P(X = 3) = \binom{3}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0 = \left(\frac{1}{2}\right)^3 = \frac{1}{8} = 0{,}125$$

$E(X) = n \times p = 3 \times 1/2 = 3/2$ e $Var(X) = n \times p \times q = 3 \times 1/2 \times 1/2 = 3/4$

A extensão para ***<u>mais do que dois eventos</u>*** (ou classes) é direta e é dada pela ***distribuição multinomial***. Se p_1 é a probabilidade associada à ocorrência do evento 1, p_2, a probabilidade do evento 2, p_3, a probabilidade do evento 3 e assim por diante, então, a probabilidade que em ***n*** ensaios independentes, o evento 1 ocorra x_1 vezes, o evento 2, x_2 vezes, o evento 3, x_3 vezes, e assim por diante, é:

$$P(x_1, x_2, x_3, ...) = \frac{n!}{x_1!x_2!x_3!...} p_1^{x_1} p_2^{x_2} p_3^{x_3} ...$$

onde: $\sum x_i = n$, $\sum p_i = 1$. Esta probabilidade é um termo na expansão de $(p_1^{x_1} + p_2^{x_2} + p_3^{x_3} + ...)^n$.

Exemplo 2. O grupo sanguíneo *MN* na população humana, onde os acasalamentos são praticamente ao acaso, apresenta os seguintes fenótipos e as respectivas probabilidades esperadas de ocorrência:

Fenótipo	Probabilidade	
MM	p^2	onde: p é a frequência do alelo M e q é
MN	$2pq$	a frequência do alelo N
NN	q^2	

Considerando uma amostra aleatória de n indivíduos dessa população, a probabilidade de que x_1 deles sejam *MM*, x_2 *MN* e x_3 *NN*, onde:

$x_1 + x_2 + x_3 = n$, é: $\frac{n!}{x_1!x_2!x_3!}(p^2)^{x_1}(2pq)^{x_2}(q^2)^{x_3}$.

4.4.3 Distribuição de Poisson

Consideremos as seguintes variáveis aleatórias:

X_1: o número de mutações num locus por geração,
X_2: o número de glóbulos vermelhos observados em cada quadrado de um hemocitômetro, e
X_3: o número de bactérias em um litro de água não-purificada,

onde: $X_i = x$, $x = 0, 1, 2, 3, ...$

O comportamento dessas variáveis aleatórias, as quais representam o número de ocorrências de eventos em um intervalo de tempo ou no espaço (*superfície ou volume*), pode ser descrito pela chamada ***distribuição de Poisson***, cuja função de probabilidade é:

$$P(X = x) = \frac{e^{-\lambda} \times \lambda^x}{x!}, \; x = 0, 1, 2, 3, ...$$

onde: $\boldsymbol{e}$ = 2,71828 e λ é o parâmetro da distribuição, que representa o número médio de ocorrências do evento por unidade de tempo ou espaço.

Uma suposição que se faz usualmente em relação a essa distribuição é que a probabilidade de se obter mais de um evento num intervalo muito pequeno é desprezível.

Esperança e Variância

Se X é uma variável aleatória com distribuição de Poisson e parâmetro λ, então, $E(X) = \lambda$ e $Var\ (X) = \lambda$. Ou seja, o número médio e a variância de ocorrências de eventos por unidade de tempo (ou espaço) são iguais (λ) e constantes ao longo do tempo (ou espaço).

Exemplo 1. Supondo que o número médio de bactérias por litro de água purificada é 2, qual é a probabilidade que 5 ou mais bactérias sejam encontradas em uma amostra de 3 litros de água?
Sendo $\lambda = 2 \times 3 = 6$, o número médio de bactérias em 3 litros de água, então:

$$P(x \geq 5) = 1 - P(X \leq 4) = 1 - \sum_{x=0}^{4} \frac{e^{-6} 6^x}{x!} = 1 - 0{,}2851 = 0{,}7149$$

Exemplo 2. Em uma população, seja X o número de descendentes produzidos por família/geração. Assumindo que $\overline{X} = \lambda = 2$, qual a probabilidade de famílias com $X = 4$ descendentes?

$$P(X=4)=\frac{e^{-2}\,.\,2^4}{4!}=0{,}0902$$

4.4.4 Distribuição de Poisson como aproximação da distribuição binomial

Algumas vezes, no uso da distribuição binomial, ocorre que ***n*** é muito grande e ***p*** é muito pequeno, de modo que ***q*** é próximo de 1. Em tais casos, o cálculo torna-se muito difícil. Pode-se, então, fazer uma aproximação da distribuição ***Binomial*** pela ***Poisson*** ou seja,

$$b(x;n,p)\cong\frac{e^{-np}(n.p)^x}{x!}$$

A aproximação é boa, se $n.p=\lambda\leq 7$.

Exemplo 1. Sabendo-se que a probabilidade de um animal ter reação negativa a certa vacina é de 0,001, determinar a probabilidade de que, de 2000 animais injetados, mais do que quatro tenham reação negativa.

$n.p=\lambda=2000\times 0{,}001=2$

$$P(X>4)=1-P(X\leq 4)=1-\left[\frac{e^{-2}2^4}{4!}+\frac{e^{-2}2^3}{3!}+\frac{e^{-2}2^2}{2!}+\frac{e^{-2}2^1}{1!}+\frac{e^{-2}2^0}{0!}\right]$$

$$=1-e^{-2}\left[\frac{16}{24}+\frac{8}{6}+\frac{4}{2}+2+1\right]=1-(0{,}135\times 7)=0{,}055$$

4.4.5 Distribuição Geométrica

Para o estudo das principais características dessa distribuição, vamos considerar uma sequência ilimitada de realizações de ensaios de Bernoulli, com probabilidade de sucesso p e probabilidade de fracasso $(1-p)=q$ em cada ensaio. Os ensaios são realizados até que ocorra o primeiro sucesso.

Nesse caso, o espaço amostral é um conjunto

S={S, FS, FFS, FFFS, FFFFS...}

Ou seja, uma sequência de n ensaios em que nos $n-1$ primeiros ensaios temos F e na n-ésima temos S.
Exemplo: para n = 8

$$\overbrace{\underbrace{F\,F\,F\,F\,F\,F\,F}_{n-1=7}\,S}^{n=8}$$

Essa distribuição apresenta duas parametrizações:

1 – O número de falhas até que ocorra o primeiro sucesso.

2 – O número de ensaios de Bernoulli necessário para obtermos o primeiro sucesso. Observe que, nesse caso, não é possível obtermos o ZERO, seu domínio será os números naturais sem o zero.

OBS: Na distribuição Binomial, o número de realizações de um ensaio era pré-determinado, enquanto aqui, na distribuição Geométrica, o número de realizações é uma variável aleatória.

Geométrica: Contagem do número de falhas até ocorrer o primeiro sucesso:

Se X a variável aleatória que fornece o número de falhas até o primeiro sucesso, essa variável tem distribuição Geométrica com parâmetro p, entre 0 e 1, e sua função é dada por:

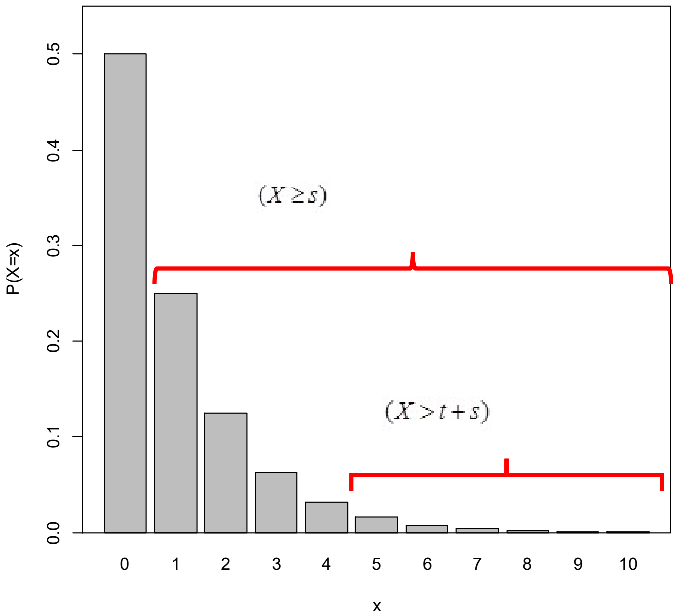
$$P(X=k) = (1-p)^k p \text{ com } k = 0, 1 \ldots$$

Usaremos a notação $X \sim \text{Geo}(p)$.
Nesse caso o evento $X = k$ ocorre, se e somente se, ocorrem somente falhas no primeiros k ensaios e sucesso no $(k + 1)$-ésimo ensaio.

Demonstração 1
Se X é uma variável aleatória discreta com distribuição geométrica, então para todo x, k = 1, 2, ... temos que:
$P(X > t + s \mid X \geq s) = P(X > t)$
Temos:

$$P(X > t + s \mid X \geq s) = \frac{P(X > t + s \cap X \geq s)}{P(X \geq s)}$$

Lembrando:
Dados os conjuntos A e B, onde:
$A \subset B$, A é um subconjunto de B, temos que:
$A \cap B = A$, assim, o conjunto $(X > t + s)$ está contido no conjunto $(X \geq s)$
Para exemplificar, dado $p= 0,5$; $t=3$ e $s = 1$



Então: $(X > t + s \cap X \geq s) = (X > t + s)$

$$P(X > t+s \mid X \geq s) = \frac{P(X > t+s)}{P(X \geq s)} = \frac{\sum_{k=t+s+1}^{\infty} p(1-p)^k}{\sum_{j=s}^{\infty} p(1-p)^j}$$

Observe que p é uma constante e pode sair das somatórias, segue que:

$$P(X > t+s \mid X \geq s) = \frac{1 - p\sum_{k=0}^{t+s}(1-p)^k}{1 - p\sum_{j=0}^{s-1}(1-p)^j} = \frac{(1-p)^{t+s+1}}{(1-p)^s} = (1-p)^{t+1} = 1 - p\sum_{k=0}^{t}(1-p)^k$$

E, portanto

$$P(X > t+s \mid X \geq s) = \sum_{k=t+1}^{\infty} p(1-p)^k = P(X > t)$$

Portanto a distribuição geométrica apresenta a propriedade de perda de memória. Vale notar que ela é a única distribuição discreta com essa característica.

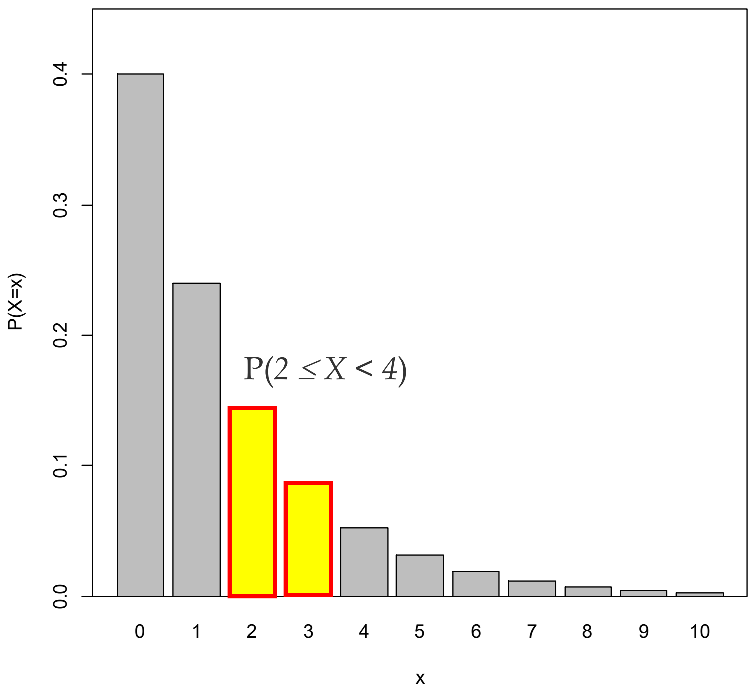
Demonstração 2

Seja $h(k) = \dfrac{P(X=k)}{P(X \geq k)}$ mostre que se X~Geo(p) então $h(k) = p$

$$h(k) = \frac{P(X=k)}{P(X \geq k)} = \frac{p(1-p)^k}{\sum_{j=k}^{\infty} p(1-p)^j} = \frac{p(1-p)^k}{1 - p\sum_{j=0}^{k-1}(1-p)^j} = \frac{p(1-p)^k}{(1-p)^k} = p$$

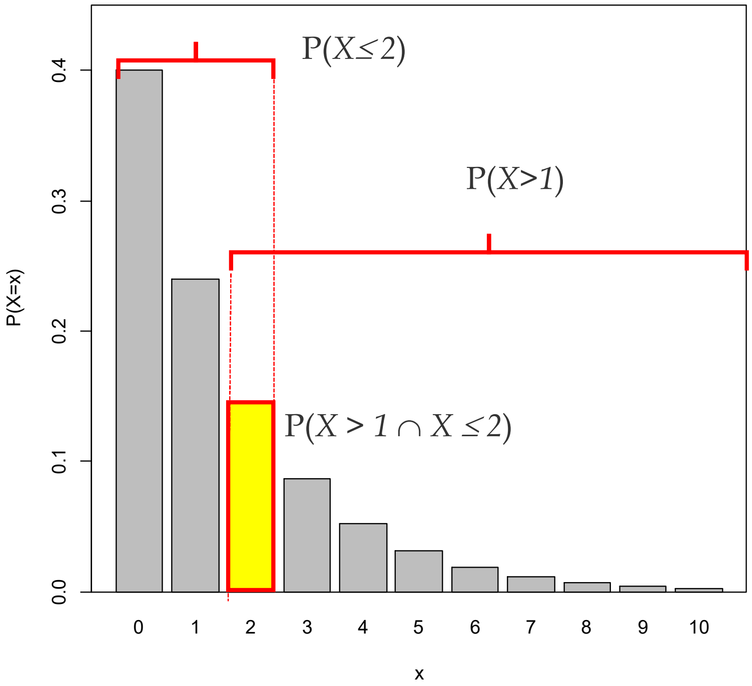
Exemplo: Considere o experimento em que uma moeda viciada é lançada sucessivas vezes, até que ocorra a primeira cara. Seja X a variável aleatória que conta o número de coroas obtidos no experimento (ou seja, a quantidade de lançamentos anteriores a obtenção da primeira cara). Sabendo que a probabilidade de cara é de 0,4, qual é a probabilidade de P($2 \leq X < 4$), a probabilidade de P($X > 1 \mid X \leq 2$) e a probabilidade de P($X \geq 1$).

Resolvendo: P($2 \leq X < 4$)



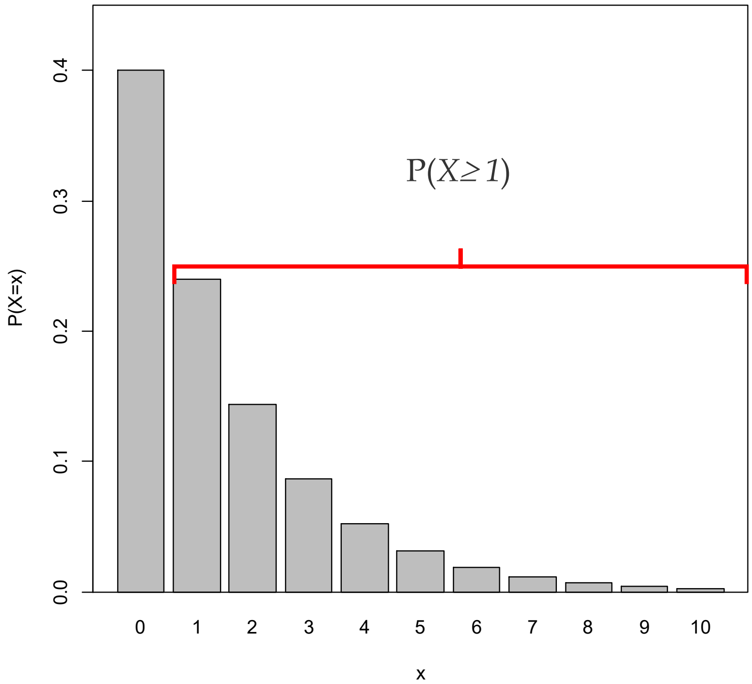
P($2 \leq X < 4$) = P(X=2) + P(X=3)= $0{,}6^2 .0{,}4 + 0{,}6^3 .0{,}4 = 0{,}2304$

Resolvendo: $P(X > 1 \mid X \leq 2)$



$$P(X > 1 \mid X \leq 2) = \frac{P(X > 1 \cap X \leq 2)}{P(X \leq 2)} = \frac{P(X = 2)}{P(X = 0) + P(X = 1) + P(X = 2)} = \frac{0{,}144}{0{,}784} = 0{,}18367$$

Resolvendo: $P(X \geq 1)$



$$P(X \geq 1) = \sum_{k=1}^{\infty} p(1-p)^k = 1 - P(X = 0) = 1 - 0{,}4.(1 - 0{,}6)^0 = 1 - 0{,}4 = 0{,}6$$

Esperança e Variância

Se X é uma variável aleatória discreta com distribuição geométrica, o valor esperado é dado por:

$$E(X)=\sum_{k=0}^{\infty}x_k\,p(x_k)=\sum_{k=0}^{\infty}k(1-p)^k\,p$$

$$E(X)=p\sum_{k=0}^{\infty}k(1-p)^k=p(1-p)\sum_{k=0}^{\infty}k(1-p)^{k-1}$$

$$E(X)=\frac{p(1-p)}{p^2}$$

$$E(X)=\frac{1-p}{p}$$

A variância será dada por:

$$Var(X)=\frac{1-p}{p^2}$$

Geométrica: Contagem do número de ensaios para se obter um sucesso:

Se X a variável aleatória que fornece o número de ensaios de Bernoulli até a obtenção do primeiro sucesso. Essa variável tem distribuição Geométrica com parâmetro p, entre 0 e 1, e sua função é dada por:

$$P(X=k)=(1-p)^{k-1}\,p,\,com\ k=1,2,...$$

A notação utilizada será X ~ Geo(p)
Observe que nessa parametrização, os valores possíveis assumidos pela variável aleatória X são R_X = {1, 2, ... } e que a probabilidade de X ser igual ao valor k é igual a probabilidade de X se igual a *k-1* na parametrização anterior.

Exemplo: Um dado honesto é lançado sucessivas vezes até que apareça pela primeira vez a face 1. Seja X a variável aleatória que conta o número de ensaios até que corra o primeiro 1.

a) Qual a probabilidade de obtermos 1 no terceiro lançamento?
b) Qual a probabilidade de obtermos 1 entre o terceiro e o 5 lançamento?

a)

$$P(X=3)=\left(1-\frac{1}{6}\right)^{3-1}.\frac{1}{6}=\frac{5^2}{6^3}=0{,}1157$$

B)

$$P(3\le X\le 5)=P(X=3)+P(X=4)+P(X=5)=\sum_{k=3}^{5}p\left(1-p\right)^{k-1}$$

$$P(3\le X\le 5)=\left(1-\frac{1}{6}\right)^{3-1}.\frac{1}{6}+\left(1-\frac{1}{6}\right)^{4-1}.\frac{1}{6}+\left(1-\frac{1}{6}\right)^{5-1}.\frac{1}{6}=\frac{25}{216}+\frac{125}{1296}+\frac{625}{7776}=0{,}29257$$

Esperança e Variância

Se X é uma variável aleatória discreta com distribuição geométrica, o valor esperado é dado por:

$$E(X) = \frac{1}{p}$$

A variância será dada por:

$$Var(X) = \frac{1-p}{p^2}$$

Exemplo: Suponha que o custo de realização de uma operação agrícola seja R$ 1.000,00. Se a operação falhar, ocorrerá um custo adicional de R$ 300,00 em virtude de serem necessárias algumas alterações antes que a próxima tentativa seja executada. Se a probabilidade de sucesso em uma tentativa for 0,2, se as operações forem independentes, e se tais operações continuarem até que a operação seja realizada com sucesso, qual será o custo esperado para esse procedimento?

Vamos definir C como o custo de operação e X a variável aleatória, ou seja, o número de tentativas necessárias para alcançar o sucesso da operação.

$$C = 1.000X + 300(X-1)$$
$$C = 1.000X + 300X - 300$$
$$C = 1.300X - 300$$

Em consequência temos:

$$E(C) = E(1.300X - 300)$$
$$E(C) = 1.300E(X) - 300$$
$$E(C) = 1.300\frac{1}{0{,}2} - 300$$
$$E(C) = R\$ \quad 6.200$$

5 VARIÁVEIS ALEATÓRIAS CONTÍNUAS

Voltemos agora nossa atenção para descrever a **distribuição de probabilidade** de uma variável aleatória (v.a.) que pode assumir todos os valores em um intervalo. Medidas de altura, temperatura, peso, produção de leite, pressão arterial, etc, são todas deste tipo.

A **distribuição de probabilidade** de uma v.a. contínua pode ser visualizada como uma forma alisada de um histograma baseado em um grande número de observações, cuja área total de todos os retângulos é igual a 1,0.

A altura do retângulo em cada intervalo de classe (Δ_i) é proporcional à densidade de proporção (f_i/Δ_i) do intervalo, de modo que a área do retângulo é igual $\Delta_i \times f_i/\Delta_i = f_i$.

Ou seja, com um número suficientemente grande de observações, diminuindo-se os intervalos de classe, o histograma tende ficar cada vez menos irregular, até aproximar da forma de uma curva bem mais suave. Isto é ilustrado na Figura 1, considerando a variável X = peso de recém-nascido.

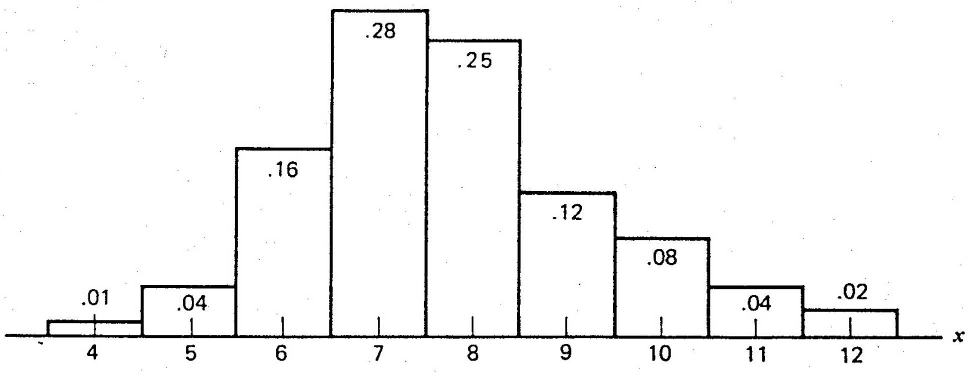
Como probabilidade é interpretada como a frequência relativa de um evento em uma longa série de ensaios independentes, a curva obtida como a forma limite dos histogramas (Figura 1c), representa a maneira pela qual a probabilidade total (1,0) é distribuída em relação à amplitude dos possíveis valores da v.a. X. A função matemática $f(x)$, cujo gráfico produz tal curva é chamada **função densidade de probabilidade** da v.a. contínua X.

A função densidade de probabilidade, $f(x)$, a qual descreve a distribuição de probabilidade para uma v.a. aleatória contínua, têm as propriedades:

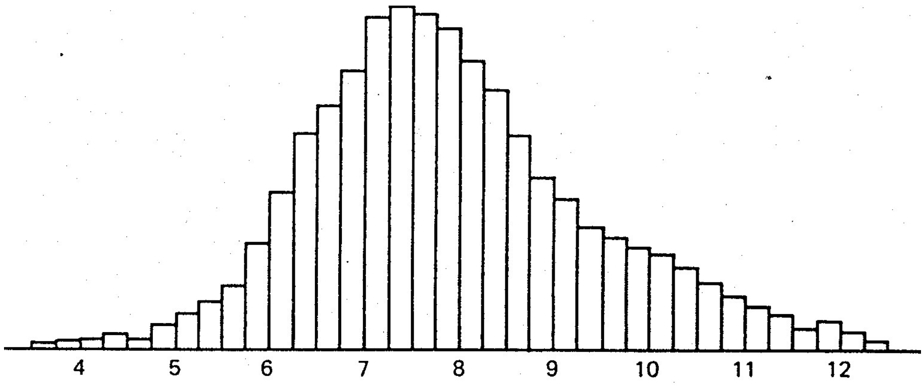
a área total sob a curva é igual a 1;
$P(a \leq X \leq b)$ = área sob a curva entre os pontos a e b;
$f(x) \geq 0$ (*não negativa*)
(d) $P(X = x_i) = 0$

"Com variáveis aleatórias contínuas, a probabilidade que $X = x_i$ é sempre zero [$P(X = x_i) = 0$]. Assim, é somente relevante falar a respeito da probabilidade que X encontra-se em um intervalo".

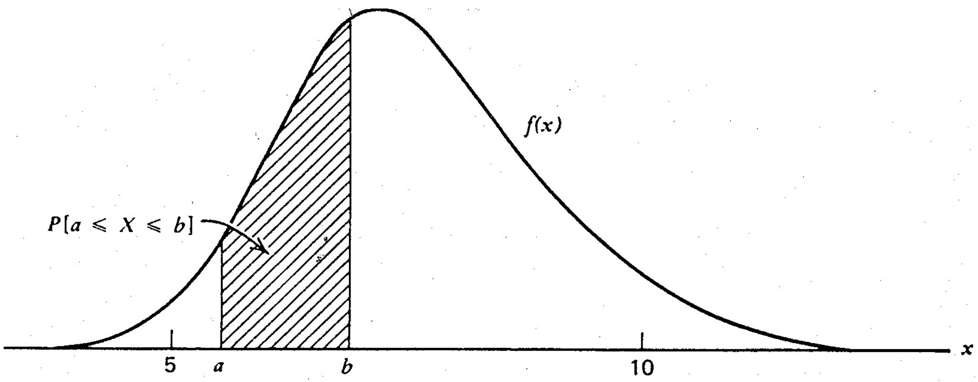
A dedução $P(X = x_i) = 0$ necessita alguns esclarecimentos. No contexto do exemplo do peso ao nascer, a afirmação $P(X = 8{,}5 \text{ lb}) = 0$, parece irreal, pois significa que nenhum recém-nascido pode pesar 8,5 lb. Para resolver este paradoxo, devemos reconhecer que a acurácia do esquema de medida é limitada, tal que o número 8,5 é indistinguível de todos aqueles que o circunda, digamos [8,495; 8,505]. Assim, a questão diz respeito à probabilidade de um intervalo circundando 8,5 e a área deste intervalo sob a curva não é maior do que zero.



Histograma de 100 pesos ao nascer com intervalos de classe de 1 libra (= 453,6g)



Histograma de 5000 pesos ao nascer com intervalos de classe de 0,25 libras.



Curva de densidade de probabilidade para a variável aleatória contínua X = peso ao nascer.

Figura 1. Curva de densidade de probabilidade vista como uma forma limite de histogramas.

Estando *f(x)* de uma variável aleatória contínua *X* especificada, o problema de se calcular $P(a \le X \le b)$, vem a ser o cálculo da área sob a curva. Tal determinação envolve cálculo integral. Mas, felizmente, áreas de distribuições importantes estão tabuladas e disponíveis para consulta.

No cálculo da probabilidade de um intervalo, *a* até *b*, não há necessidade de se preocupar se qualquer um dos extremos ou ambos estão incluídos no intervalo. Com $P(X = a) = P(X = b) = 0$,

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

Valem para as v.a contínuas, os conceitos de ***esperança*** (μ) e ***variância*** (σ^2). Suas determinações, entretanto, exigem a aplicação de método de cálculo integral que não será aqui utilizado.

Dada uma v.a. X contínua, interessa saber qual a $f(x)$. Alguns modelos são frequentemente usados para representar a função densidade de probabilidade (*f.d.p.*) de v.a. contínuas. O mais utilizado é descrito a seguir:

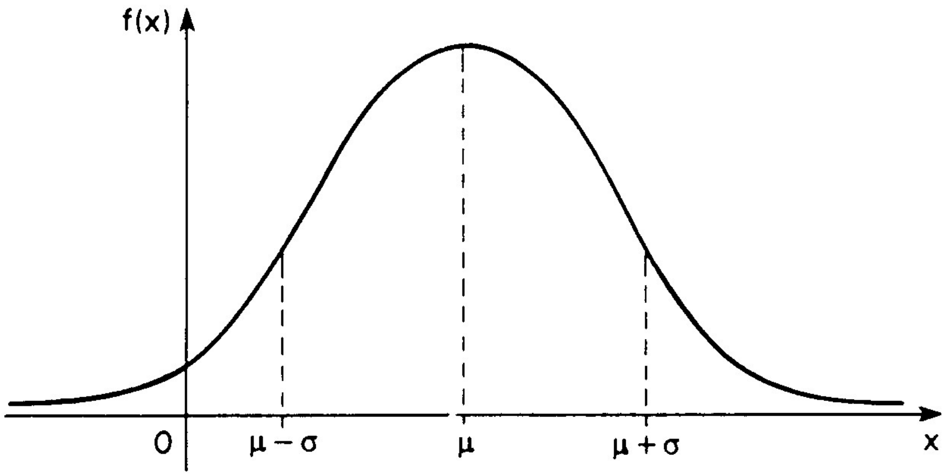
5.1 Distribuição Normal

Definição: Uma v.a. X tem distribuição normal com parâmetros μ e σ^2, $-\infty < \mu < \infty$ e $0 < \sigma^2 < \infty$, se sua *f.d.p.* é dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \ -\infty < x < \infty \qquad (1)$$

onde: $\pi = 3{,}14159...$; $e = 2{,}71828$ ...

Gráfico



5.1.1 Propriedades

Os parâmetros μ e σ^2 representam, respectivamente, a média e a variância da distribuição, isto é, $E(X) = \mu$ e $Var(X) = \sigma^2$. A demonstração requer manipulações de integral e não será apresentada aqui.

Outras propriedades, enumeradas a seguir, podem ser facilmente observadas de seu gráfico:

$f(x) \rightarrow 0$ quando $x \rightarrow \pm\infty$

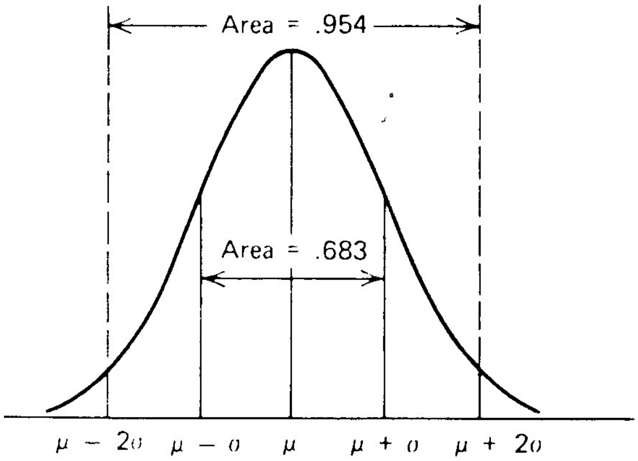
$\mu - \sigma$ e $\mu + \sigma$ são pontos de inflexão de $f(x)$

$x = \mu$ é o ponto de máximo de $f(x)$ e o valor máximo é $\dfrac{1}{\sigma\sqrt{2\pi}}$

$f(x)$ é simétrica ao redor de $x = \mu$, isto é, $f(\mu + x) = f(\mu - x)$, para todo $-\infty < x < \infty$

média = moda = mediana

Os intervalos $\mu \pm \sigma$, $\mu \pm 2\sigma$ e $\mu \pm 3\sigma$, têm, respectivamente, as probabilidades de 0, 683, 0,954 e 0,997, ou seja:

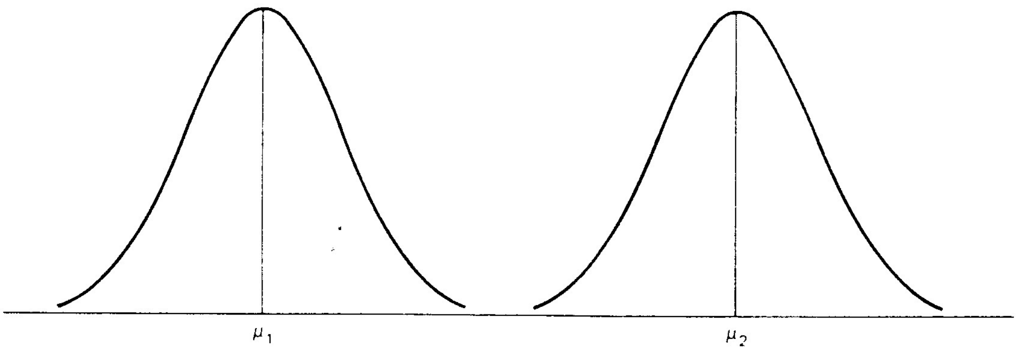


Distribuição normal

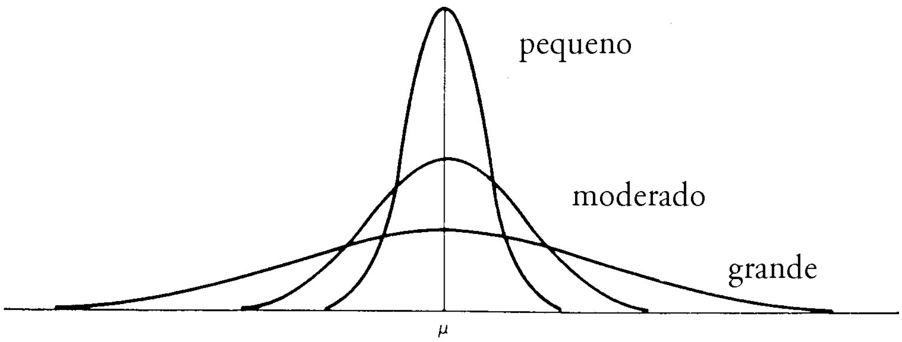
Se X tem distribuição normal, com média μ e variância σ^2, denota-se por:

$\boldsymbol{X : N(\mu, \sigma^2)}$

Interpretando os parâmetros

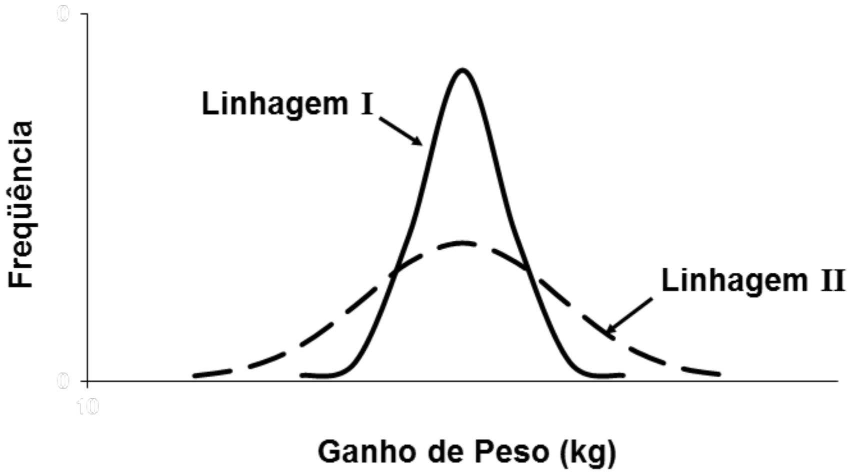


Duas distribuições normais com diferentes médias, mas com o mesmo desvio padrão (σ)



Três distribuições normais com médias iguais, mas com diferentes desvios padrões (σ). Decrescendo σ, aumenta a altura máxima $(1/\sigma\sqrt{2\pi})$ e a concentração de probabilidade em torno de μ.

Exemplo 1. Considere dois grupos de frangos de corte criados em uma granja no sul de Minas Gerais, comparáveis em todos os aspectos, exceto pela linhagem.



O gráfico ilustra o ganho de peso dessas populações e permite afirmar que:

( ) a média aritmética e a variância da Linhagem I são superiores às da Linhagem II.
( ) a média aritmética da Linhagem I é superior à da II e as variâncias são iguais.
( ) as médias aritméticas são iguais e a variância da Linhagem I é superior à da II.
(x) as médias aritméticas são iguais e a variância da Linhagem I é inferior à da II.
( ) a média aritmética e a variância da Linhagem I são inferiores às da Linhagem II.

5.2 Distribuição normal padronizada

A distribuição dada por (1) representa uma família de distribuições, dependendo dos valores μ e σ^2. A particular distribuição normal com $\mu = 0$ e $\sigma^2 = 1$ é referida como **distribuição normal padronizada ou reduzida**. Sua média e variância coincidem com as da variável

$$Z = \frac{X - \mu}{\sigma} \quad (2)$$

onde $X : N(\mu, \sigma^2)$

A variável Z é chamada **variável normal padronizada**, cuja função densidade pode ser obtida de (1), fazendo-se formalmente $\mu = 0$ e $\sigma = 1$, isto é:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \quad (-\infty < z < \infty)$$

Se *X: N*(μ, σ^2), então a variável aleatória *Z* definida por (2) terá uma distribuição *N(0, 1)*. Mostraremos que *Z* tem média 0 e variância 1:

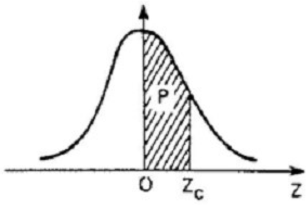
$$E(Z) = E(\frac{X}{\sigma} - \frac{\mu}{\sigma}) = E(\frac{X}{\sigma}) - \frac{\mu}{\sigma} = \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$$

$$Var(Z) = E(Z^2) - [E(Z)]^2 = E[(\frac{X-\mu}{\sigma})^2] = \frac{1}{\sigma^2}E(X-\mu)^2 = \frac{1}{\sigma^2}\sigma^2 = 1$$

mas, não é fácil é mostrar que *Z* tem distribuição normal e não será demonstrado aqui.
A curva normal padrão, *f(z)*, é também simétrica em torno de μ_z e as áreas sob a curva nos intervalos de –1 a +1(μ ± σ), –2 a +2 (μ ± 2σ) e –3 a +3 (μ ± 3σ), são também iguais a, respectivamente, 68,3%, 95,4% e 99,7% da área total, que é 1.
A vantagem de se usar a variável Z é que as áreas, ou as probabilidades, associadas à distribuição normal padronizada são tabeladas (ver Tabela 2). Assim, a transformação (2) é fundamental para o cálculo de probabilidades relativas a uma distribuição normal qualquer.

Tabela 2

Distribuição normal reduzida: N(0;1)
Probabilidades p tais que $p = P(0 < Z < Z_c)$

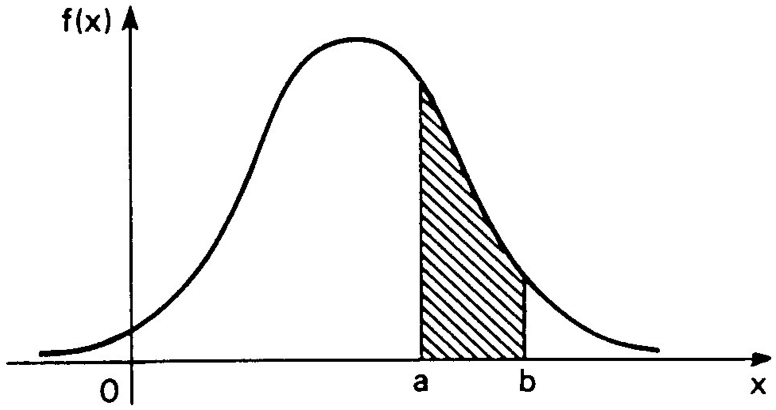


parte inteira e primeira decimal de Z_c	SEGUNDA DECIMAL DE Z_c 0	1	2	3	4	5	6	7	8	9	parte inteira e primeira decimal de Z_c
	p = 0										
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586	0,0
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535	0,1
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409	0,2
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173	0,3
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793	0,4
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240	0,5
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490	0,6
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524	0,7
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327	0,8
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891	0,9
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214	1,0
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298	1,1
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147	1,2
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774	1,3
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189	1,4
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408	1,5
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449	1,6
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327	1,7
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062	1,8
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670	1,9
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169	2,0
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574	2,1
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899	2,2
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158	2,3
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361	2,4
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520	2,5
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643	2,6
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736	2,7
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807	2,8
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861	2,9
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900	3,0
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929	3,1
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950	3,2
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965	3,3
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976	3,4
3,5	49977	49978	49978	49979	49980	49981	49981	49982	49983	49983	3,5
3,6	49984	49985	49985	49986	49986	49987	49987	49988	49988	49989	3,6
3,7	49989	49990	49990	49990	49991	49991	49992	49992	49992	49992	3,7
3,8	49993	49993	49993	49994	49994	49994	49994	49995	49995	49995	3,8
3,9	49995	49995	49996	49996	49996	49996	49996	49996	49997	49997	3,9
4,0	49997	49997	49997	49997	49997	49997	49998	49998	49998	49998	4,0
4,5	49999	50000	50000	50000	50000	50000	50000	50000	50000	50000	4,5

parte inteira e primeira decimal de Z_c	SEGUNDA E TERCEIRA DECIMAIS DE Z_c 05	15	25	35	45	55	65	75	85	95	parte inteira e primeira decimal de Z_c
	p = 0,										
0,0	00199	00598	09997	01396	01795	02193	02591	02989	03387	03784	0,0
0,1	04181	04578	04974	05369	05764	06159	06553	06946	07339	07730	0,1
0,2	08121	08512	08901	09290	09677	10064	10450	10834	11218	11600	0,2
0,3	11982	12362	12741	13119	13495	13871	14244	14617	14988	15358	0,3
0,4	15726	16093	16458	16822	17184	17545	17903	18261	18500	18970	0,4
0,5	19322	19672	20021	20368	20712	21055	21396	21735	22073	22408	0,5
0,6	22741	23072	23401	23729	24054	24377	24697	25016	25333	25647	0,6
0,7	25959	26270	26577	26883	27186	27488	27786	28083	28377	28669	0,7
0,8	28959	29246	29531	29814	30094	30372	30648	30921	31192	31461	0,8
0,9	31727	31990	32252	32511	32767	33021	33273	33522	33769	34013	0,9

Aplicação

Suponha que X : $N(\mu, \sigma^2)$ e queiramos determinar ***P(a < X < b)***, tal como representado na figura a seguir:



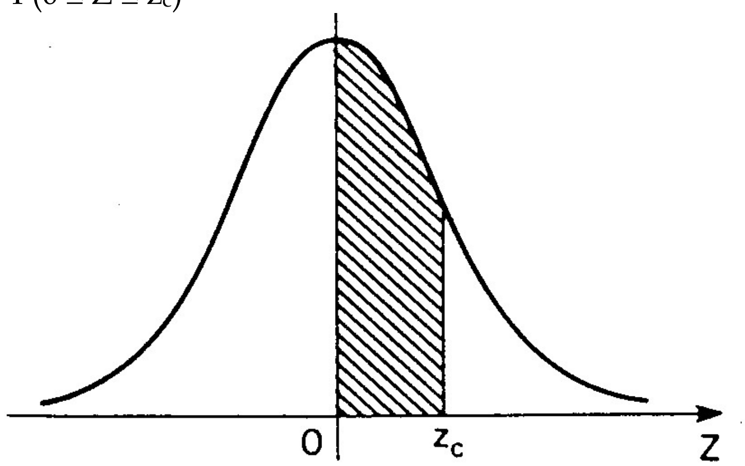
Por exemplo, tomando $a = 2$ e $b = 5$ e supondo que $X: N(3, 16)$, calculemos

$P(2 \leq X \leq 5)$

Vejamos, antes, como obter probabilidades a partir da Tabela 2 para a distribuição $N(0,1)$.

A figura abaixo ilustra a probabilidade fornecida pela tabela, ou seja,

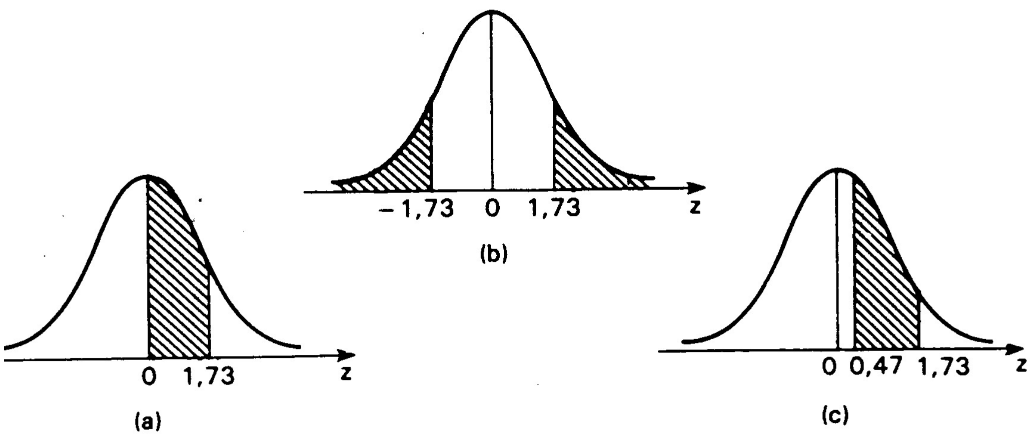
$P(0 \leq Z \leq z_c)$



Se $z_c = 1{,}73$

$P(0 \leq Z \leq 1{,}73) = 0{,}4582$

Observe:



$P(-1{,}73 \leq Z \leq 0) = P(0 \leq Z \leq 1{,}73) = 0{,}4582$, devido à simetria da curva

$P(Z \geq 1{,}73) = 0{,}5 - P(0 \leq Z \leq 1{,}73) = 0{,}5 - 0{,}4582 = 0{,}0418$

$P(Z < -1{,}73) = P(Z > 1{,}73) = 0{,}0418$

$P(Z \leq 1{,}73) = P(Z \geq -1{,}73) = P(0 \leq Z \leq 1{,}73) + P(Z < 0) = 0{,}4582 + 0{,}5 = 0{,}9582$

$P(0{,}47 \leq Z \leq 1{,}73) = P(0 \leq Z \leq 1{,}73) - P(0 \leq Z \leq 0{,}47) = 0{,}4582 - 0{,}1808 =$

$= 0{,}2774$

Para usar a Tabela 2 em conexão com uma variável aleatória X, tendo distribuição normal, deve-se efetuar a mudança de escala $Z = \dfrac{X-\mu}{\sigma}$. Assim, no exemplo,

$$P(2 \leq X \leq 5) = P\left(\frac{2-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{5-\mu}{\sigma}\right)$$

$$= P\left(\frac{2-3}{4} \leq Z \leq \frac{5-3}{4}\right) = P(-1/4 \leq Z \leq 1/2)$$

Pela tabela N(0,1):

$P(-0{,}25 \leq Z \leq 0{,}5) = P(-0{,}25 \leq Z \leq 0) + P(0 < Z \leq 0{,}5)$

$P(-0{,}25 \leq Z \leq 0{,}5) = 0{,}0987 + 0{,}1915 = 0{,}2902$ ou seja,

$P(2 \leq x \leq 5) = 0{,}2902$

Exemplo 1. Sabendo-se que os pesos à desmama (X) de 10.000 bezerros de um rebanho são distribuídos normalmente, com média (μ) 170 kg e desvio padrão (σ) 5 kg, (a) qual é o número esperado de bezerros com peso superior a 165 kg?; e (b) que peso (x) deve atingir um bezerro para que ele supere 80% dos pesos à desmama desse rebanho?

Solução:

(a) P(X > 165) = $P\left(\frac{X-\mu}{\sigma} > \frac{165-170}{5}\right) = P(Z > -1)$

P(Z > -1) = P(-1< Z ≤ 0) + P(Z > 0) =

= 0,3413 + 0,5 = 0,8413

Portanto, o número esperado é 10.000 × 0,8413 ≅ 8.413 bezerros.

(b) P(X ≤ 170) **+** P(170 **<** X < x) = 0,80

0,5 + P(170 **<** X < x) = 0,80

P(170 **<** X < x) = 0,30 e P(X ≥ x) = 0,20

P(170 **<** X < x) = $P\left(0 < \frac{X-\mu}{\sigma} < \frac{x-170}{5}\right) = \mathrm{P}\left(0 < Z < \frac{x-170}{5}\right) = 0{,}30$ e

P(X ≥ x) = 0,5 - $P\left(0 < Z < \frac{x-170}{5}\right) = 0{,}20$

$z_c = \frac{x-170}{5} = 0{,}84 \Rightarrow$ x = 174,2kg

5.3 Aproximação Normal à Binomial

Se X tem distribuição binomial b(n, p), onde n é grande e p não é muito próximo de 0 ou 1, a distribuição da variável padronizada $Z = \frac{X-np}{\sqrt{np(1-p)}}$ é aproximadamente N(0,1). Assim,

$$P(a \le X \le b) = \sum_{x=a}^{b} \binom{n}{x} p^x (1-p)^{n-x} \cong P\left[\frac{a-np}{\sqrt{np(1-p)}} \le Z \le \frac{b-np}{\sqrt{np(1-p)}}\right]$$

$$P(a \le X \le b) \cong P\left[\frac{a-np}{\sqrt{np(1-p)}} \le Z \le \frac{b-np}{\sqrt{np(1-p)}}\right]$$

Tendo em vista que uma distribuição discreta (binomial) é aproximada por uma contínua (normal), a melhor aproximação é obtida calculando:

$$P(a \le X \le b) \cong P\left[\frac{(a-0{,}5)-np}{\sqrt{np(1-p)}} \le Z \le \frac{(b+0{,}5)-np}{\sqrt{np(1-p)}}\right]$$

Dividindo-se os numeradores e denominadores do intervalo de Z por ***n***, pode-se também escrever:

$$P(a \leq X \leq b) \cong P\left[\frac{[(a-0,5)/n]-p}{\sqrt{[p(1-p]/n}} \leq Z \leq \frac{[(a+0,5)/n]-p}{\sqrt{[p(1-p)]/n}}\right]$$

O termo **±1/(2n)** é chamado ***"correção de continuidade"***.

Exemplo 2. Supondo que X : b(15, 0,4)

$$P(7 \leq X \leq 10) = \sum_{x=7}^{10}\binom{15}{x}0,4^{x}(0,6)^{15-x} = 0,381$$

$$P(7 \leq X \leq 10) \cong P\left(\frac{7-6}{1,9} \leq Z \leq \frac{10-6}{1,9}\right)$$

$$\cong P\left(0,526 \leq Z \leq 2,105\right) = 0,48257 - 0,20194 = 0,281$$

Usando **correção de continuidade**:

$$P(7 \leq X \leq 10) \cong P\left(\frac{6,5-6}{1,9} \leq Z \leq \frac{10,5-6}{1,9}\right)$$

$$\cong P(0,263 \leq Z \leq 2,368) = 0,49111 - 0,10194 = 0,389$$

Para justificar a correção de continuidade, basta atentar para a Figura 2.

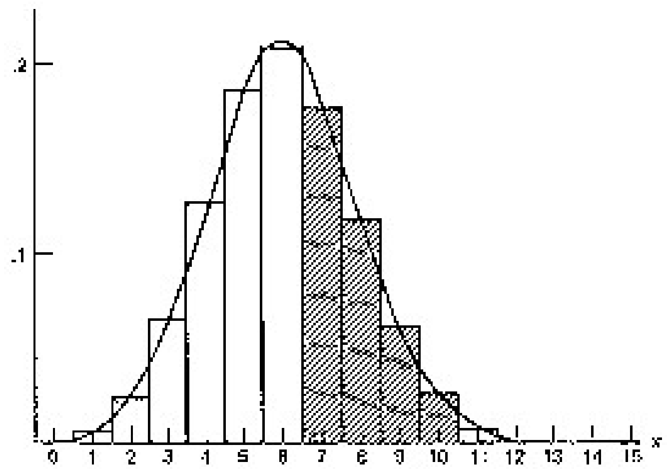


Figura 2. Histograma da distribuição binomial b(15, 0,4) e a curva normal aproximada.

A distribuição normal pode ser recomendada para aproximar probabilidades binomiais, mesmo para ***n*** tão pequeno quanto 15, contanto que ***p*** seja próximo de 1/2. Quando ***p*** é muito pequeno e ***n*** é grande, a distribuição de ***Poisson*** é mais apropriada. Como uma norma prática, ***n*** pode ser assumido como "suficientemente" grande para se usar a distribuição normal, quando:

$np(1-p) \geq 3$

sendo que a aproximação melhora com o crescimento de *n*.

6 AMOSTRAGEM

Na realização de qualquer estudo quase nunca é possível examinar todos os elementos da população de interesse, seja por questão de tempo ou econômica. Outras vezes, a análise é destrutiva, por exemplo, de vacinas, remédios, etc. Assim, a solução é selecionar parte dos elementos (amostra), analisá-la e inferir propriedades para o todo (população). Este é o objetivo da **Inferência Estatística**. Dois conceitos básicos são necessários para o desenvolvimento da Inferência Estatística: **população** e **amostra**.

População é o conjunto de indivíduos (objetos), tendo pelo menos uma variável comum observável.

Amostra é qualquer subconjunto da população.

No momento em que decidimos obter informações por meio de um levantamento amostral, temos de imediato definir a população de interesse e selecionar a característica que iremos estudar. A **população-alvo** é a população sobre a qual iremos fazer inferências baseadas na amostra.

A maneira de se obter a amostra é tão importante, e existem tantos modos de fazê-lo, que estes procedimentos constituem uma especialidade dentro da Estatística, conhecida como **Amostragem**. Tais procedimentos podem ser agrupados em dois grupos: os chamados **planos probabilísticos** e **planos não probabilísticos**.

O primeiro grupo reúne as técnicas que usam mecanismos aleatórios de seleção dos elementos da amostra, atribuindo a cada um deles uma probabilidade, conhecida ***a priori***, de fazer parte da amostra. Mais especificamente, dizemos que um método de seleção produz amostras probabilísticas, se ele define claramente a probabilidade de um dado elemento vir a fazer parte da amostra.

No segundo grupo estão os demais procedimentos, tais como: **amostras intencionais** ou de **"peritos"**, onde os elementos são selecionados com auxílio de especialistas e **amostras de conveniência**, onde o critério para a seleção dos elementos é dado pela facilidade de acesso a esses elementos. Muitas vezes as amostras de conveniência são constituídas por **voluntários**, como ocorre em testes sobre a eficiência de vacinas.

Para que possamos fazer inferências válidas sobre uma população a partir de uma única amostra dela extraída, é preciso que esta seja representativa da população. Uma das formas de se conseguir representatividade é fazer com que o processo de escolha da amostra seja, de alguma forma aleatório, isto é, de modo casual. Além disso, a aleatoriedade permite o cálculo de estimativas dos erros envolvidos no processo de inferência. Estas são as razões pelas quais as amostras probabilísticas são preferidas.

Descreveremos a seguir os métodos mais comuns de extração de amostras probabilísticas. Ao descrevê-los, estaremos sempre tratando de obter uma amostra de tamanho n em uma população de tamanho N.

6.1 Amostragem aleatória simples ou amostragem aleatória sem reposição

Amostragem aleatória simples ou amostragem aleatória sem reposição é o delineamento amostral no qual, n distintos elementos são selecionados de N elementos na população, de tal maneira, que cada combinação possível de r elementos, é igualmente provável ser a amostra selecionada. A amostra pode ser obtida por r seleções em que, em cada passo, todos os elementos não selecionados da população, têm igual chance de seleção. Equivalentemente, pode-se tomar uma sequência de seleções independentes da população total, tendo cada elemento, em cada passo, igual probabilidade de seleção, descartando seleções repetidas e continuando até que ***r*** elementos distintos sejam obtidos.

Com este tipo de amostragem, a probabilidade que o *i*-ésimo elemento da população seja incluído na amostra é $p_i = r/N$, de modo que a probabilidade de inclusão é a mesma para cada elemento. Outros delineamentos podem atribuir a cada elemento igual probabilidade de ser incluído na amostra, mas somente com amostra aleatória simples, cada possível amostra de ***r*** elementos tem a mesma probabilidade de ocorrência.

Uma amostra aleatória simples pode ser selecionada escrevendo os elementos da população, numerados de **1** a ***N***, em ***N*** cartões, misturando-os numa urna e sorteando, sem reposição, ***r*** desses cartões. Ou seja, a amostra consiste daqueles elementos da população, cujas identificações correspondem aos números selecionados. Existirão $\binom{N}{r} = \dfrac{N!}{r!(N-r)!}$ amostras possíveis diferentes de tamanho ***n***.

Leitura	26	04	28	30	22
Peso	29	19	28	29	31

Pode-se usar um procedimento alternativo, escolhendo-se numa tábua de números aleatórios ou usando algoritmos computacionais que geram números aleatórios, ***n*** números compreendidos entre **1** e ***N***. Os elementos correspondentes aos números escolhidos formarão a amostra. Evidentemente, devem ser desprezados números já escolhidos (já estão na amostra).

Tábuas de números aleatórios são coleções de dígitos construídos aleatoriamente e que simulam o processo de sorteio. A Tabela 3 apresenta um pequeno conjunto de tais números.

Exemplo 1. A tabela a seguir refere-se aos pesos (kg) ao nascer de 30 bezerros da raça Gir de uma fazenda (dados hipotéticos).

Bezerro	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**
Peso	26	32	26	19	20	22	30	31	25	20	27
Bezerro	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Peso	28	28	27	26	19	23	25	25	26	27	31
Bezerro	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**			
Peso	21	26	23	29	30	28	24	29			

Extrair, sem reposição, uma amostra aleatória de tamanho ***n*** = 5.

Solução. Lendo uma coluna da Tábua I, digamos a primeira, tomamos os 5 primeiros números não superiores a 30. Obtemos, assim, a amostra:

Leitura	25	12	22	07	11
Peso	23	28	31	30	27

Poderíamos, também, escolher a terceira coluna. Obteríamos a amostra:

6.2 Amostragem aleatória simples com reposição

Imaginemos agora que os elementos da amostra (***r***) são selecionados um de cada vez, a partir dos elementos da população (***N***), repondo o elemento sorteado na população antes do próximo sorteio. Com tal procedimento, qualquer elemento pode ser sorteado mais do que uma vez. Uma amostra de elementos assim selecionados é chamada ***amostra aleatória simples com reposição***. As ***r*** seleções são independentes e cada elemento na população tem a mesma probabilidade de inclusão na amostra. Amostra aleatória com reposição é caracterizada pela propriedade que cada possível sequência de ***r*** unidades, distinguindo ordem de seleção e possibilidade de inclusão de seleções repetidas, tem igual probabilidade sob o delineamento amostral.

Uma vantagem prática deste tipo de amostragem é que, em algumas situações, é uma conveniência importante não ser necessário averiguar se qualquer elemento nos dados está incluído na amostra mais de uma vez. Entretanto, para um dado tamanho amostral ***r***, a ***amostra aleatória simples com reposição***, como será visto no próximo capítulo, é menos eficiente do que ***a sem reposição*** para estimar o valor médio (μ) de uma população.

6.3 Amostragem aleatória estratificada

Quando os elementos da população estão divididos em grupos distintos, é mais fácil e eficiente escolher, independentemente, uma ***amostra aleatória simples*** dentro de cada um desses grupos, os quais são chamados ***estratos***.

Esta forma de amostragem é uma das mais utilizadas, já que a maioria das populações têm estratos bem definidos. Como exemplo, imagine que se deseje obter uma amostra de vacas em lactação responsáveis pelo abastecimento de leite de uma usina de beneficiamento. Deve ser considerado que esta é constituída por distintos rebanhos (estratos) fornecedores.

Então, para obter uma amostra de vacas em lactação que seja mais representativa da usina, deve-se selecionar uma amostra dentro de cada estrato, isto é, uma amostra dentro de cada rebanho, e depois reunir as amostras em uma só, constituindo assim uma **amostra estratificada**.

O mais comum é utilizar a **amostragem estratificada proporcional**, que consiste em selecionar os elementos da amostra entre os vários estratos, em número proporcional ao tamanho de cada um dos estratos. Deste modo, sendo:

N - o número de elementos da população
L - o número de estratos
N_i - o número de elementos do estrato i
n - o tamanho da amostra a ser selecionada,

onde:

$$N = N_1 + N_2 + ... + N_L$$

calcula-se a fração de amostragem por $f = \frac{n}{N}$, e o número de elementos a serem sorteados em cada estrato será:

$N_1.f, \quad N_2.f, ..., N_L.f$

Exemplo 2. Supondo que se deseje estimar a taxa de ocorrência de mastite sub-clínica em vacas em lactação que abastecem a usina de beneficiamento, extrair, sem reposição, uma amostra estratificada de tamanho **n** = 8, considerando que há dois rebanhos fornecedores: A e B, respectivamente, com 10 e 35 vacas em lactação.

Solução. No rebanho A as vacas são numeradas de 1 a 10 e no B de 1 a 35. A fração de amostragem é:

$$f = \frac{8}{45} = 0{,}18$$

De cada estrato (rebanho) serão sorteados respectivamente n_A e n_B elementos (vacas):

$n_A = 0{,}18.10 = 1{,}8 \cong 2$

$n_B = 0{,}18.35 = 6{,}3 \cong 6$

Escolhendo uma coluna da Tábua I, digamos a segunda, obtemos o resultado:

Estrato	A		B					
Leitura	09	01	09	01	06	15	35	12

Extraída a amostra, a taxa de ocorrência de mastite sub-clínica é estimada pesquisando a ocorrência da doença na mesma.

Dentre as vantagens da amostra estratificada destacam-se:

a) Os dados são geralmente mais homogêneos dentro de cada estrato do que na população como um todo;

b) Podem-se obter estimativas separadas dos parâmetros populacionais para cada estrato sem selecionar outra amostra e, portanto, sem custo adicional;

c) Na amostragem casual simples, as unidades amostradas podem não cobrir todos os elementos da população, principalmente quando **n** é muito menor do que **N**. Então, a amostragem estratificada é mais eficiente e preferível à aleatória simples.

6.4 Amostragem por conglomerado

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

O primeiro passo para se usar esse processo é especificar conglomerados apropriados. Os elementos em um conglomerado devem ter características semelhantes. Como regra geral, o número de elementos em um conglomerado deve ser pequeno em relação ao tamanho da população e o número de conglomerados, razoavelmente grande.

Tanto na amostragem estratificada, como na amostragem por conglomerado, a população deve estar dividida em grupos. Na amostragem estratificada, entretanto, seleciona-se uma **amostra aleatória simples dentro de cada grupo (estrato),** enquanto que na amostragem por conglomerado selecionam-se **amostras aleatórias simples de grupos**, e todos os elementos dentro dos grupos (conglomerados) selecionados farão parte da amostra.

A amostragem por conglomerado é recomendada quando:

a) Ou não se tem um sistema de referência listando todos os elementos da população, ou a obtenção dessa listagem é dispendiosa;

b) O custo da obtenção de informações cresce com o aumento da distância entre os elementos.

Exemplo 3. Supondo agora que se deseje estimar a taxa de ocorrência de mastite sub-clínica em vacas em lactação considerando várias usinas de beneficiamento, como deve ser escolhida a amostra?

Solução. A **amostragem aleatória simples** é inviável, pois pressupõe uma listagem de todas as vacas em lactação que abastecem as usinas, o que é muito difícil de se obter.

A alternativa da **amostragem estratificada** é também inviável, já que aqui também é necessária uma listagem dos elementos por estrato (rebanho).

A melhor escolha é a **amostragem por conglomerado**. O sistema de referência pode ser constituído por todos os rebanhos fornecedores de leite às usinas. Cada rebanho é um conglomerado. Extrai-se uma **amostra aleatória simples** de rebanhos e neles pesquisa-se a ocorrência de mastite em todas as vacas em lactação.

6.5 Amostragem sistemática

Neste processo de amostragem, os elementos são selecionados para a amostra por um sistema pré-estabelecido, que seja completamente alheio à natureza da variável em estudo. Assim, uma **amostra sistemática** de tamanho **n** pode ser constituída, como uma sugestão, dos elementos de ordem

k, k + r, k + 2r, k + 3r, ...

onde: **k** é um número inteiro escolhido aleatoriamente entre 1 e **n** e **r** é o inteiro mais próximo da fração **N/n**. Por exemplo, se a população tem 100 elementos (**N** = 100) e vamos escolher uma amostra de tamanho 6 (**n** = 6), **k** é um inteiro escolhido aleatoriamente entre 1 e 6 e **r** = $\frac{100}{6} = 16,6 \cong 17$. Se **k** = 3, a amostra será composta pelos seguintes elementos:

3 20 37 54 71 88

Se o tamanho da população é desconhecido, não podemos determinar exatamente o valor de **r**. Escolheremos intuitivamente um valor razoável para **r**.

Nos casos em que a população está organizada, a **amostragem sistemática** é preferível à **amostragem aleatória simples**, porque é mais fácil de executar, estando, portanto, menos sujeita a erros.

Exemplo 4. Vamos supor que um pesquisador pretenda obter uma amostra de prontuários veterinários para estudar a proporção de cães internados devido à cinomose. Se o número do prontuário é conferido por ordem de chegada do animal no hospital e é razoável pressupor que a ordem de chegada independa do motivo de internamento, o pesquisador pode obter uma amostra sistemática selecionando todos os prontuários cujos números terminam em determinados dígitos, digamos 2. Assim, a amostra será constituída de prontuários de ordem 2, 12, 22, 32, ... , o que corresponde a **k** = 2 e **r** = 10, de acordo com o esquema anterior.

7 ESTATÍSTICA E DISTRIBUIÇÃO AMOSTRAL

A estatística se interessa por conclusões e predições originadas de resultados eventuais que ocorrem em experimentos ou investigações cuidadosamente planejados.
Esses resultados eventuais constituem um subconjunto ou **amostra** de medidas ou observações de um conjunto maior de valores, chamado **população**. No entanto, nem todas as amostras prestam para validar generalizações a respeito de populações, das quais foram obtidas. Muitos dos métodos de inferência são baseados em **amostras aleatórias simples com reposição**.

7.1 Amostra aleatória simples com reposição

Definição 1. Uma **amostra aleatória simples com reposição** de tamanho **n** de uma variável aleatória *X* com uma dada distribuição é o conjunto de **n** variáveis aleatórias independentes *X1*, *X2*, ..., *Xn*, cada uma com a mesma distribuição de *X*. Assim, por exemplo, se *X* tem distribuição *b(n, p)*, cada X_i terá distribuição *b(n, p)*.

7.2 Estatísticas e parâmetros

Definição 2. **Estatística** ou estimador é qualquer função de uma amostra aleatória (fórmula ou expressão), construída com o propósito de servir como instrumento para descrever alguma característica da amostra e para fazer inferência a respeito da característica na população. A(o)s mais comuns são:

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$: média da amostra

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}(\sum_{i-1}^{n} X_i^2 - \overline{X}^2)$: variância da amostra

$\hat{p} = \frac{X}{n} = \frac{(\text{número de elementos da amostra que apresentam a característica})}{(\text{tamanho da amostra})}$: proporção da amostra

Definição 3. Parâmetro **é uma medida usada para descrever uma característica da população.**

Parâmetros são funções de valores ***populacionais***, enquanto que ***estatísticas*** são funções de valores ***amostrais***. Os símbolos mais comuns são:

Estatística	População
Média: $\overline{X}$,	$E(X) = \mu$
Variância: s^2	σ^2
Nº de elementos: n	N
Proporção: $\hat{p}$	p

7.3 Distribuição amostral

Toda estatística, sendo uma função de uma amostra aleatória $X_1, X_2, ..., X_n$, é também uma variável aleatória e tem uma distribuição. Embora, em uma dada situação estaremos limitados apenas a uma amostra e um valor único correspondente à estatística; em relação a várias amostras, a estatística muda de valor de acordo com a distribuição determinada a partir daquela que controla a amostra aleatória. O ponto importante é que o comportamento da estatística pode ser descrito por alguma distribuição de probabilidade. Assim, cada estatística é uma variável aleatória e sua distribuição de probabilidade é chamada ***distribuição amostral da estatística***. Esquematicamente, teríamos o procedimento apresentado na Figura 1, onde θ é o parâmetro de interesse na população e t é o valor da estatística T para cada amostra.

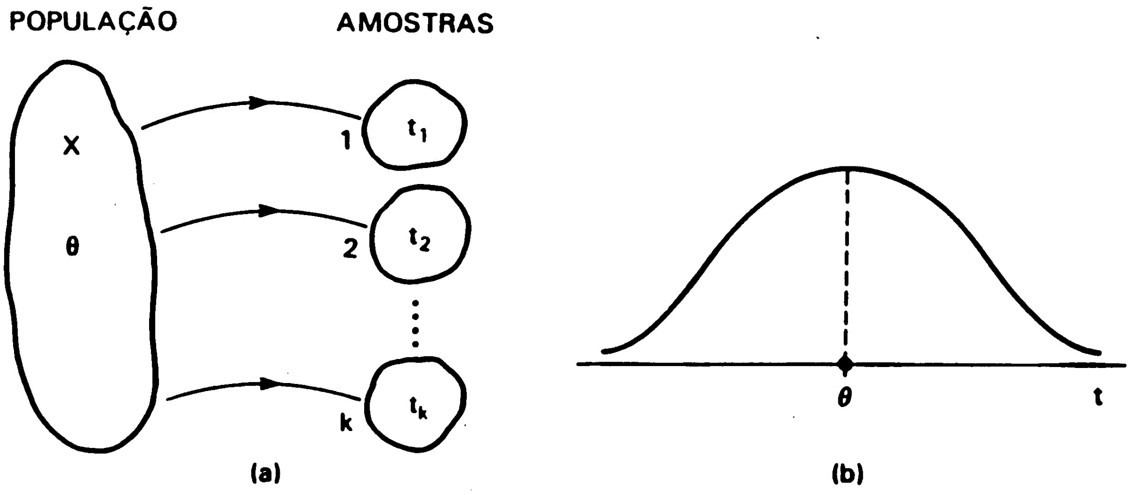


Figura 1: (a) amostras retiradas da população, de acordo com certo procedimento, e (b) distribuição amostral da estatística T.

O exemplo abaixo ilustra como a distribuição da média amostral pode ser determinada por uma situação simples, quando o tamanho da amostra é 2 (n = 2) e a distribuição da população é discreta.

Exemplo1. Seja a variável aleatória X que denota o número de dias de internação de um cão em um hospital veterinário depois de uma particular cirurgia. Considerando a população de todos os cães submetidos à cirurgia, suponha que X tem a distribuição de probabilidade apresentada na Tabela 1. Uma amostra aleatória simples com reposição (X_1, X_2) de 2 cães (n = 2) é tomada nesta população. Qual a distribuição do número médio amostral de dias de internação, ou seja:

$$\overline{X} = \frac{X_1 + X_2}{2} = ?$$

Tabela 1. Distribuição de probabilidade de X

x	0	1	2	3
$p(x)$	0,2	0,4	0,3	0,1

De acordo com a definição de amostra aleatória simples com reposição, X_1 e X_2 são variáveis aleatórias independentes, cada uma tendo a distribuição dada na Tabela 1. Deste modo, a distribuição conjunta de duas variáveis aleatórias independentes (Tabela 2) é obtida multiplicando-se as probabilidades marginais. Por exemplo:

$P(X_1 = 0, X_2 = 1) = P(X_1 = 0)\times P(X_2 = 1) = 0{,}2\times 0{,}4 = 0{,}08$

A distribuição de $\overline{X}$ é obtida por meio da Tabela 2, listando os possíveis valores de $\overline{X}$. Em seguida, para cada valor de $\overline{X}$, identificamos as células na referida tabela, cujos valores (X_1, X_2) produzem um específico valor de $\overline{X}$. Então, somamos as correspondentes probabilidades celulares. Por exemplo:
$\overline{X} = 1{,}5$ quando $(X_1, X_2) = (0, 3), (1, 2), (2, 1)$ ou $(3, 0)$, tal que $P[\overline{X} = 1{,}5] = 0{,}02 + 0{,}12 + 0{,}12 + 0{,}02 = 0{,}28$. Procedendo de modo análogo, obtemos a distribuição amostral da estatística $\overline{X}$ (Tabela 3).

Tabela 2. Distribuição conjunta de X_1 e X_2:

x_1	x_2: 0	1	2	3	Σ linha
0	0,04	0,08	0,06	0,02	0,20
1	0,08	0,16	0,12	0,04	0,40
2	0,06	0,12	0,09	0,03	0,30
3	0,02	0,04	0,03	0,01	0,10
Σ coluna	0,20	0,40	0,30	0,10	1,00

Tabela 3. Distribuição amostral de $\overline{X} = \dfrac{X_1 + X_2}{2}$:

Valor de $\overline{X}$	0	0,5	1	1,5	2	2,5	3	Total
Probabilidade	0,04	0,16	0,28	0,28	0,17	0,06		1,0
	0,01							

7.4 Distribuição amostral da média e o teorema limite central

Resultados importantes:

1. Se X_1, X_2, ..., X_n constitui uma amostra aleatória simples com reposição de uma população que tem média μ e variância σ^2, então:

$$E(\overline{X}) = \mu \text{ e } Var(\overline{X}) = \frac{\sigma^2}{n}$$

Verifiquemos essas relações, considerando a variável aleatória discreta exemplificada (Exemplo 1):

Distribuição de X:

x	0	1	2	3	Total
$p(x)$	0,2	0,4	0,3	0,1	1,0
$x \times p(x)$		0	0,4	0,6	1,3
$x^2 \times p(x)$	0	0,4	1,2	0,9	2,5

$\mu = E(X) = \sum x \times p(x) = 1{,}3$

$\sigma^2 = E(X^2) - [E(X)]^2 = \sum x^2 \times p(x) - \left[\sum x \times p(x)\right]^2$

$\sigma^2 = 2{,}5 - (1{,}3)^2 = 0{,}81$

Distribuição de $\overline{X} = \dfrac{X_1 + X_2}{2}$:

$\bar{x}$	0	0,5	1	1,5	2	2,5	3	Total
$p(\bar{x})$	0,04	0,16	0,28	0,28	0,17	0,06	0,01	1,0
$\bar{x} \times p(\bar{x})$	0	0,08	0,28	0,42	0,34	0,15	0,03	1,3
$\bar{x}^2 \times p(\bar{x})$	0	0,04	0,28	0,63	0,68	0,375	0,09	2,095

$E(\overline{X}) = \sum \bar{x} \times p(\bar{x}) = 1{,}3 = E(X)$

$Var(\overline{X}) = E(\overline{X})^2 - [E(\overline{X})]^2 = \sum \bar{x}^2 \times p(\bar{x}) - [E(\overline{X})]^2$

$Var(\overline{X}) = 2{,}095 - (1{,}3)^2 = 0{,}405 = \dfrac{\sigma^2}{n} = \dfrac{0{,}81}{2}$

Assim, a distribuição da ***média amostral***, baseada em uma amostra aleatória simples com reposição de tamanho ***n***, tem:

$E(\overline{X}) = \mu$ (média da população)

$Var(\overline{X}) = \dfrac{\sigma^2}{n}$ (variância da população)

$dp(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$ (desvio padrão da população / $\sqrt{n}$) = **erro padrão da média**

O desvio padrão da média$[dp(\overline{X})]$ e o erro padrão da média são termos equivalentes. **O erro padrão da média** é geralmente usado para evitar confusão com o desvio padrão (σ) das observações.

Esses resultados mostram que a distribuição da média amostral ($\overline{X}$) é centrada na média populacional μ e que o cálculo de $\overline{X}$ produz uma estatística que é menos variável do que uma observação individual (*X*). Com o aumento do tamanho da amostra (*n*), o desvio padrão (dp) da distribuição de $\overline{X}$ diminui. Isto significa que quando *n* torna-se grande, podem-se esperar valores de $\overline{X}$ mais próximos de μ, a quantidade que se pretende estimar.

Normalmente não se tem várias amostras para se obter estimativas múltiplas da média. No entanto, é possível estimar o erro padrão da média usando o tamanho da amostra (*n*) e desvio padrão (*s*) de uma única amostra de observações. O erro padrão da média é, então, estimado pelo desvio padrão das observações dividido pela raiz quadrada do tamanho da amostra.

À medida que o tamanho da amostra aumenta, o desvio padrão da amostra (*s*) irá flutuar, mas não vai aumentar ou diminuir de forma consistente. Torna-se uma estimativa mais precisa do desvio padrão paramétrico (σ) da população. Em contraste, o erro padrão da média torna-se menor quando o tamanho da amostra aumenta. Com tamanhos amostrais maiores, a média da amostra torna-se uma estimativa mais precisa da média paramétrica (μ), pois o erro padrão da média torna-se menor.

Os resultados precedentes são principalmente de interesse teórico. De valor prático maior são dois outros resultados, que serão mencionados a seguir, sem demonstrá-los:

2. Se $\overline{X}$ é a média de uma amostra aleatória simples com reposição, de tamanho ***n***, de uma população ***normal***, com média μ e variância σ^2, sua distribuição é ***normal***, com média μ e variância $\dfrac{\sigma^2}{n}$.

O outro é o **teorema limite central** (ou teorema central do limite):

3. Em uma amostra aleatória simples com reposição de uma população **arbitrária**, com média μ e variância σ^2, a distribuição de $\overline{X}$, quando ***n*** é ***grande***, é aproximadamente normal, com média μ e variância $\dfrac{\sigma^2}{n}$. Em outras palavras,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \text{ é aproximadamente N (0,1)}$$

Uma ilustração gráfica do teorema limite central aparece na Figura 2, onde a distribuição da população representada pela curva contínua é uma distribuição contínua assimétrica, com μ = 2 e σ = 1,41. As distribuições da média amostral $\overline{X}$ para tamanhos amostrais n = 3 e n = 10 são representadas no gráfico pelas curvas pontilhadas, indicando que, com um aumento de n, as distribuições amostrais tornam-se mais concentradas ao redor de μ, assemelhando-se a uma **distribuição normal**.

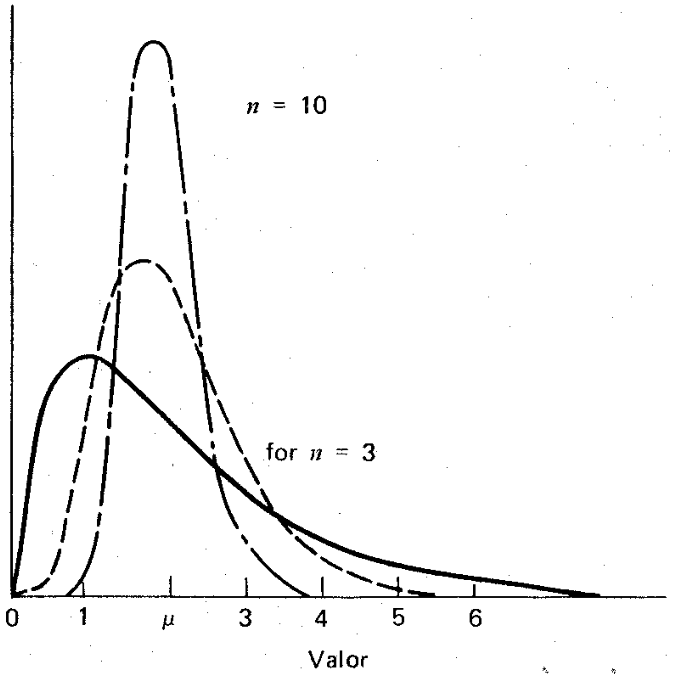


Figura 2. Distribuições de $\overline{X}$ para n = 3 e n = 10 amostradas em uma população com distribuição assimétrica (curva contínua).

Na prática, a aproximação é usada quando $n \geq 30$**, indiferente da forma da população amostrada.**

Aplicação do teorema limite central

O teorema limite central tem muitos aspectos práticos úteis: se $\overline{X}$ é a média amostral, podemos calcular:

$$P\,(a < \overline{X} \leq b) = P\,\left(\frac{a-\mu}{\sigma/\sqrt{n}} < Z \leq \frac{b-\mu}{\sigma/\sqrt{n}}\right)$$

aproximadamente, usando tabelas da distribuição *N(0,1)*, qualquer que seja a distribuição de X.

As distribuições de outras estatísticas, por exemplo, da proporção amostral $\hat{p}$ (veja item 3.2), também podem ser aproximadas pela distribuição normal, assumindo ***n*** grande.

Exemplo 2. Seja uma máquina de empacotamento de um determinado sal mineral, cujos pesos (em kg) seguem uma distribuição N(50, 2). Assim, se a máquina estiver regulada, qual a probabilidade, colhendo-se uma amostra de 100 pacotes, da média dessa amostra diferir de 50 kg em menos de 0,2828 kg?

Solução:

$$P\,(\,49{,}7172 < \overline{X} < 50{,}2828\,) = P\,(\,\frac{49{,}7172-50}{\sqrt{2}/10} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{50{,}2828-50}{\sqrt{2}/10})$$

$= P\,(\,-2{,}0 < Z < 2{,}0\,)$

$= 2 \times P\,(0 < Z < 2{,}0) = 2 \times 0{,}47725$

$= 0{,}9545$

Ou seja, dificilmente 100 pacotes terão uma média fora do intervalo]49,7172; 50,2828[. Caso apresentem uma média fora desse intervalo, pode-se considerar como sendo um evento raro, e será razoável desconfiar que a máquina esteja desregulada.

Amostras sem reposição de populações finitas

Supondo uma população com **N** elementos, se a amostragem for feita **sem reposição**, $\mathbf{E(\overline{X}) =} \ \mu$ continua a valer, mas $\mathbf{Var(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}}$, em que ${}^{N-n}\!/\!_{N-1}$ é o **fator de correção para população finita**.

A variância da média amostral com este tipo de amostragem é menor do que com reposição, pois ela é igual a ${}^{N-n}\!/\!_{N-1}$ vezes a variância da média amostral, quando a amostragem é com reposição (σ^2/n). Disto se deduz que a amostragem sem reposição é mais eficiente do que a com reposição para estimar o valor médio (μ). No entanto, se a população for grande quando comparada com o tamanho da amostra (**n**), o fator de correção será próximo de 1 e $\mathrm{Var}(\overline{X}) \approx {}^{\sigma^2}\!/\!_{n}$, consequentemente, a diferença na eficiência entre o dois tipos de amostragens torna-se desprezível. Esta aproximação pode ser usada, se $\mathbf{n} \le \mathbf{5\%\ N}$.

Note que quando **n** se aproxima de **N**, o fator de correção se aproxima de zero, de modo que a $\mathbf{Var(\overline{X})}$ também se aproxima de zero.

7.5 Distribuição amostral da proporção

Designemos uma variável *X* para cada ensaio de ***Bernoulli***, onde há somente dois resultados possíveis: **Sucesso (S)** e **Fracasso (F)**, com $P(S) = p$. Neste contexto, considerando ***n*** ensaios independentes, $X_1, X_2, \ldots, X_n$ constitui uma amostra aleatória simples com reposição. Como os resultados individuais são 0 (fracasso) ou 1 (sucesso), $\sum_{i=1}^{n} X_i$ é o número de resultados em ***n*** ensaios, que correspondem aos sucessos (ou ao número de elementos amostrados que possuem uma específica característica), porque aos resultados que correspondem aos fracassos, estão associados o valor zero. Então,

$T = X_1 + X_2 + ... + X_n = \sum_{i=1}^{n} X_i$ = número de sucessos em ***n*** ensaios. Portanto, a proporção amostral de sucessos é $\hat{p} = \frac{T}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$ ou seja, $\hat{p}$ é igual à média da variável aleatória X_i $(i = 1, 2, ..., n)$.

T tem distribuição binomial $b(n, p)$, com média np e variância npq. Consequentemente,

$$E(\hat{p}) = E\left(\frac{T}{n}\right) = \frac{1}{n} E(T) = \frac{1}{n}\, n\, p = p$$

$$Var(\hat{p}) = Var\left(\frac{T}{n}\right) = \frac{1}{n^2} Var(T) = \frac{1}{n^2}\, n\, p\, q = \frac{p\, q}{n}$$

Assim, pelo **Teorema Limite Central**, quando **n** é **grande**, a proporção amostral $\hat{p}$ de sucessos em n ensaios de Bernoulli tem distribuição aproximadamente normal com média p e variância $\frac{p\, q}{n}$; e

$Z = \frac{\hat{p} - p}{\sqrt{p\, q / n}}$ é aproximadamente $N\ (0,\ 1)$

Multiplicando-se o numerador e o denominador de Z por ***n*** e notando-se que $n\,\hat{p} = T$, pode-se também escrever

$Z = \frac{T - n\, p}{\sqrt{n\, p\, q}} \sim N\ (0,\ 1)$,

que foi o estabelecido na aproximação normal à binomial.

Exemplo 3. Um lote 625 vacas foram inseminadas com sêmen que possui índice de fertilidade (p) de 70%. Qual a probabilidade de se encontrar mais de 72% (450) de vacas prenhes?

Solução:

$n = 625$ $p = 0{,}70$

$$P(\hat{p} > 0{,}72) \cong P(Z > \frac{0{,}72 - 0{,}70}{\sqrt{\frac{0{,}70 \times 0{,}30}{625}}}) \cong P(Z > 1{,}09) \cong 0{,}50 - 0{,}36214 \cong 0{,}1379$$

Ou $P(T > 450) \cong P(Z > \frac{450 - 437{,}5}{\sqrt{0{,}7.0{,}30.625}}) \cong P(Z > 1{,}09) \cong 0{,}1379$

7.6 Estimação de uma proporção binomial

Consideremos os tipos de problemas, onde o parâmetro é a proporção ***p*** de uma população, tendo uma específica característica. Quando ***n*** elementos são aleatoriamente amostrados da população, os dados consistirão da contagem ***X*** do número de elementos amostrados possuindo a característica. O senso comum sugere a proporção amostral:

$$\hat{p} = X/n$$

como um estimador de ***p***. Quando ***n*** é uma pequena fração do tamanho da população, como geralmente é o caso, observações à respeito de ***n*** elementos podem ser consideradas como sendo de ***n*** ensaios independentes de Bernoulli, com probabilidade de sucesso igual a ***p***.

Quanto às propriedades desse estimador, primeiro nota-se que a contagem amostral ***X*** tem distribuição binomial ***b(n, p)***, com média ***np*** e variância ***npq***, onde ***q = 1 – p***. Consequentemente,

$$E(\hat{p}) = E(\tfrac{X}{n}) = \tfrac{1}{n}E(X) = \frac{np}{n} = p$$

$$Var(\hat{p}) = Var(\tfrac{X}{n}) = \tfrac{1}{n^2}Var(X) = \frac{npq}{n^2} = \frac{pq}{n}$$

O primeiro resultado mostra que $\hat{p}$ é um estimador não viciado de p. O segundo, que $\hat{p}$ tem uma variância que é menor do que a variância de qualquer outro estimador não viciado. O erro padrão desse estimador é dado por:

$$dp(\hat{p}) = \sqrt{\frac{pq}{n}}$$

o qual pode ser obtido substituindo ***p*** e ***q*** pelas suas respectivas estimativas amostrais, ou seja $\hat{p}$ e $\hat{q}$, na fórmula, ou $dp(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$

Assim, como foi observado no item anterior, quando ***n*** é grande, $\hat{p}$ é aproximadamente distribuído como normal, com média ***p*** e desvio padrão $\sqrt{\frac{pq}{n}}$; e $Z = \frac{\hat{p} - p}{\sqrt{p\,q/n}}$ é aproximadamente N (0, 1).

8 ESTIMAÇÃO

A maioria dos trabalhos em estatística é realizada com o uso de **amostras aleatórias** extraídas de uma população, na qual se deseja fazer um determinado estudo.
A parte da estatística que procura deduzir informações relativas a uma população, mediante a utilização de amostras dela extraídas, é denominada **Inferência Estatística**.
Um dos problemas da estatística é a estimativa de parâmetros populacionais (média, variância, proporção, etc), mediante o uso de uma estatística amostral (média amostral, variância amostral, proporção amostral, etc).

Definição. O valor numérico da **estatística** ou **estimador** de um parâmetro, calculado para uma amostra observada, é chamado de **estimativa** desse parâmetro.
A diferença entre estatística e estimativa é que a estatística é uma variável aleatória, e a estimativa é um particular valor dessa variável aleatória.

8.1 Propriedades de um bom estimador

8.1.1 Consistência

Consistência é uma propriedade por meio da qual a acurácia de uma estimativa aumenta quando o tamanho da amostra aumenta.
Um estimador $(\hat{\theta})$ é chamado consistente se a probabilidade dele diferir do verdadeiro valor θ em menos do que c, onde c é um número arbitrário positivo e pequeno, tende a 1, quando o tamanho da amostra (n) aumenta; ou seja, se

$$\lim_{n \to \infty} P(\left| \hat{\theta} - \theta \right|) = 1$$

Isto significa que, quando ***n*** aumenta, a estimativa $\hat{\theta}$ torna-se mais provável estar próxima (dentro de uma distância fixada pequena, ± c) do verdadeiro parâmetro θ. Isto é uma propriedade assintótica de um estimador. Ela é aplicada a amostras "suficientemente grandes". As condições suficientes para um estimador ser consistente são:

$\lim_{n \to \infty} E(\hat{\theta}) = \theta$ e $\lim_{n \to \infty} Var(\hat{\theta}) = 0$

Vejamos um exemplo para ilustrar. Considere a distribuição amostral da média, baseada em amostras aleatórias simples com reposição de tamanho n; obtém-se $E(\overline{X}) = \mu$ *e* $Var(\overline{X}) = \dfrac{Var(X)}{n}$. À medida que ***n*** cresce a distribuição de $\overline{X}$ torna-se mais concentrada em torno de μ. Diz-se que $\overline{X}$ é um estimador consistente da média da população (μ). Do mesmo modo, o estimador $\hat{p}$ é tal que $Var(\hat{p}) \to 0$, quando $n \to \infty$; chamamo-lo de consistente devido a este fato e a que $E(\hat{p}) = p$.

8.1.2 Não viciado ou não viesado

Um estimador, $\hat{\theta}$, como uma variável aleatória, tem uma certa distribuição em repetidas amostras de tamanho ***n***. Em uma particular amostra, o valor calculado pode desviar em mais ou menos de θ, mas espera-se que, em média, ele determine o verdadeiro valor (θ). Não viciado é uma propriedade que assegura que, em média, o estimador é correto.

O estimador $\hat{\theta}$ é chamado não viciado ou imparcial se seu valor esperado ou médio for igual ao verdadeiro valor do parâmetro, θ, isto é, $E(\hat{\theta}) = \theta$. Qualquer estimador $\hat{\theta}$, para o qual $E(\hat{\theta}) = \theta + b(\theta)$, com $b(\theta) \neq 0$, é chamado viciado; a quantidade $b(\theta)$ é chamada vício ou viés.

Por analogia com experimentos químicos ou bioquímicos, o vício corresponde ao "erro sistemático" ou "erro do método". Um químico pode usar um certo método para o qual os resultados obtidos, em experimentos repetidos, podem ser muito próximos um do outro, mas, em média, não dão a resposta correta. Situação similar pode ocorrer com um estatístico na construção de um estimador. Todavia, nem sempre é necessário preocupar-se em obter um estimador não viciado, pois quando o tamanho da amostra aumenta, o $\lim_{n\to\infty} E(\hat{\theta}) = \theta$, tal que $\hat{\theta}$ é assintoticamente não viciado.

Exemplos. Como foi mostrado, $E(\overline{X}) = \mu$, isto é, $\overline{X}$ é um estimador não viciado da média da população (μ) e $E(\hat{p}) = p$, ou seja, $\hat{p}$ é um estimador não viciado de p. Estes estimadores nada mais são do que as próprias definições dos respectivos parâmetros, mas aplicadas à amostra.

Por outro lado, o estimador da variância da população $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2$, dado por $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, é viciado, pois, como pode ser demonstrado, $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{1}{n}\sigma^2$, onde $b(\sigma^2) = -\frac{1}{n}\sigma^2$. Tomando-se o estimador "ajustado" $\frac{n}{n-1}\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, então s^2 é um estimador não viciado para σ^2, porque $E(s^2)$ $= E\left(\frac{n}{n-1}\hat{\sigma}^2\right) = \frac{n}{n-1}E(\hat{\sigma}^2) = \sigma^2$. Por esta razão, s^2 foi definido como a variância amostral. No entanto, para $n \to \infty$, têm-se para ambos os estimadores: $\lim_{n\to\infty} E(\hat{\sigma}^2) = \lim_{n\to\infty} E(s^2) = \sigma^2$, isto é, $\hat{\sigma}^2$ *e* s^2 são assintoticamente não viciados.

Deve ser mencionado que, embora s^2 seja um estimador não viciado da variância σ^2, s não é um estimador não viciado do desvio padrão σ. Também pode ser mostrado que um estimador não viciado da covariância entre duas variáveis X e Y, é a covariância amostral:

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

8.2 Estimativa por ponto e por intervalo

A estimativa de um parâmetro populacional dada por um único valor para a estatística é denominada **estimativa por ponto**. Por exemplo, a estimativa pontual da média populacional μ é feita por um valor $\overline{X}$. Todavia, esse procedimento não permite julgar qual a possível magnitude do erro que se está cometendo. Daí surge a idéia de construir os **intervalos de confiança**, que são baseados na distribuição amostral do estimador pontual.

A estimativa de um parâmetro populacional dada por dois valores **a** e **b** (a < b), entre os quais se considera que o parâmetro esteja contido, é denominada **estimativa por intervalo**.

As estimativas por intervalo indicam a sua precisão ou exatidão, por isto são preferíveis às estimativas por ponto. A declaração da precisão de uma estimativa por intervalo denomina-se **grau de confiança** ou **nível de confiança**. Daí a denominação de **Intervalo de Confiança**.

Exemplo 1. Dizendo-se que o diâmetro da artéria aorta em bovinos tem uma medida de 1,75 cm, está-se apresentando uma estimativa por ponto. Por outro lado, se for dito que o diâmetro mede 1,75 ± 0,05 cm, a estimativa é por intervalo, isto é, afirma-se que o diâmetro da aorta está entre 1,70 e 1,80 cm.

8.3 Estimativas por intervalos de confiança

Formalmente, seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de tamanho **n** e θ um parâmetro desconhecido da população. Um intervalo de confiança para θ é um intervalo construído a partir das observações da amostra, de modo que ele inclui o verdadeiro e desconhecido valor de θ, com uma específica e alta probabilidade. Esta probabilidade, denotada por **1 - α**, é tipicamente tomada como 0,90; 0,95 ou 0,99. Indica-se por:

$P(a < \theta < b) = 1 - \alpha$

Então, o intervalo **] a, b [** é chamado **intervalo com 100 (1 - α)% de confiança** para o parâmetro θ, onde: **1 - α** é o **nível de confiança** associado ao intervalo e **a** e **b** são os **limites de confiança,** inferior e superior, respectivamente, do intervalo.

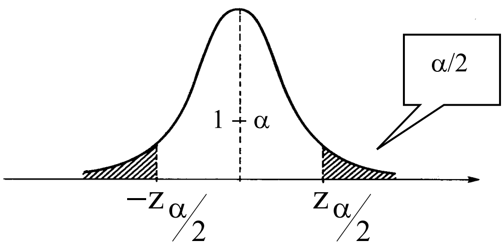
8.3.1 Para a média populacional (μ)

(a) Caso em que n é grande e σ conhecido.

O desenvolvimento de intervalos de confiança para μ é baseado na distribuição amostral de $\bar{X}$. Sabe-se que, pelo **Teorema Limite Central**, se o tamanho da amostra (n) é grande, $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ é aproximadamente $N(0,1)$.

Usando-se a tabela da distribuição $N(0,1)$, pode-se determinar um valor $z_{\alpha/2}$, tal que :

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$



$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(-\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

onde:

$$\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} = a \quad \text{e} \quad \bar{X} + z_{\alpha/2}\sigma/\sqrt{n} = b$$

Denomina-se: $\begin{cases} \dfrac{\sigma}{\sqrt{n}} = \sigma_{\overline{X}} = \textit{erro padrão da média} \\ z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}} = \textit{erro da estimativa da média} \end{cases}$

Se $1 - \alpha = 0{,}95$

$$P(\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95$$

Esta expressão deve ser interpretada do seguinte modo: construídos todos os intervalos da forma $\overline{X} \pm 1{,}96\,\sigma_{\overline{X}}$, 95% deles conterão μ (veja Figura 1). Lembrando que μ não é uma variável aleatória, mas um parâmetro, isto não é o mesmo que dizer que μ tem 95% de probabilidade de estar entre os limites indicados.

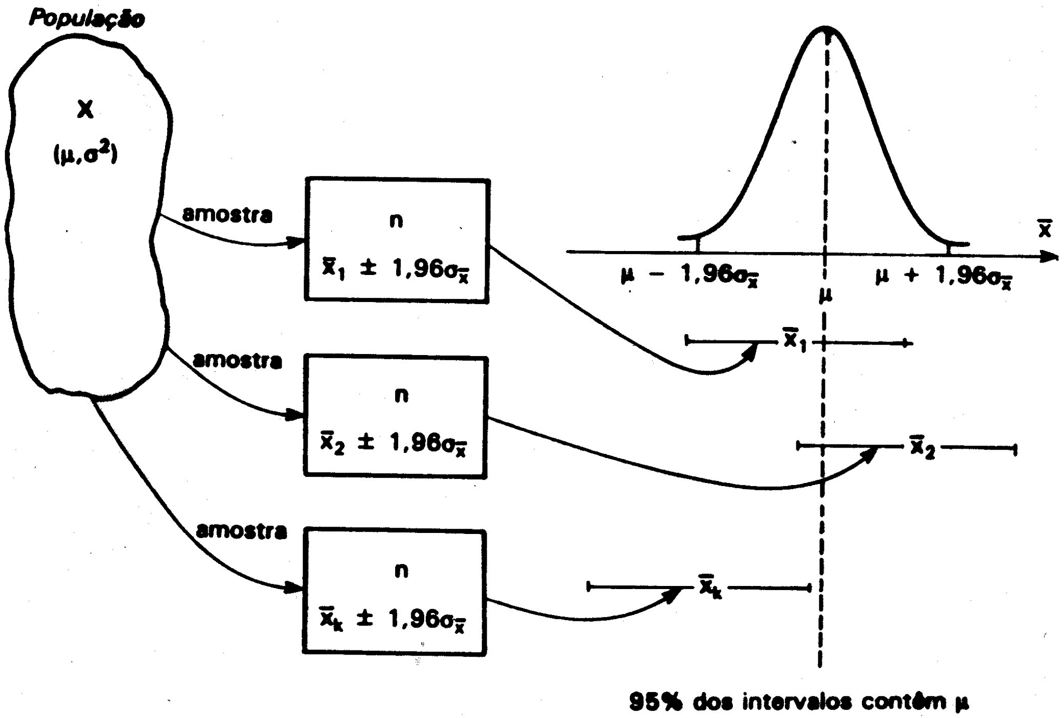


Figura 1. Significado de um IC para μ, com $(1 - \alpha) = 0{,}95$ e σ^2 conhecido

Selecionada uma amostra, encontrada sua média ($\bar{x}_a$) e sendo conhecido $\sigma_{\bar{x}}$, pode-se construir o intervalo:

$$\bar{x}_a \pm 1{,}96\sigma_{\bar{x}}$$

Este intervalo pode ou não conter o parâmetro μ, mas, pelo exposto acima, têm-se 95% de confiança de que o contenha.

Indica-se um intervalo de 100 $(1 - \alpha)$% de confiança para μ, quando n é grande e σ conhecido, por:

$$IC(\mu : 1-\alpha) =]\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ;\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}[$$

Se $(1 - \alpha) = 0{,}95 \rightarrow z_{\alpha/2} = 1{,}96$

Em um intervalo com:

nível de confiança (1 - α) fixo, se o tamanho da amostra (n) aumenta, a amplitude do intervalo $(A = 2 \,.\, z_{\alpha/2} \,.\, \sigma/\sqrt{n})$ diminui;

n fixo, se (1 - α) aumenta, A também aumenta, pois o valor de $z_{\alpha/2}$ aumenta.

Exemplo 2. Considerando uma amostra de 100 animais da raça Nelore, onde o peso médio a desmama é 171,70 kg, encontre um IC de 95% para μ, supondo que o desvio padrão da população (σ) seja igual a 7,79 kg.

Solução:

$$\text{IC}(\mu : 95\%) = 171{,}70\,\text{kg} \pm 1{,}96.\frac{7{,}79\,\text{kg}}{\sqrt{100}} =]170{,}17\,\text{kg}\,;173{,}23\,\text{kg}[$$

(b) Caso em que n é grande e σ desconhecido

Para grandes amostras, a afirmação probabilística

$$P(\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} < \mu < \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha$$

é ainda correta, mas como σ é desconhecido, o intervalo não pode ser construído. Entretanto, como **<u>n é grande (n ≥ 30), a substituição de σ pelo desvio padrão amostral (s) não afeta apreciavelmente essa afirmação probabilística, pois o valor numérico de s é uma estimativa acurada de σ</u>**, de modo que $Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ é aproximadamente *N(0,1)*. Assim, o IC(μ : 1– α) é dado por:

$$\left]\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}\ ; \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right[$$

(c) Para a média populacional μ com base em amostras pequenas (n < 30)

Se $X_1, X_2, ..., X_n$ é uma amostra aleatória de uma população com distribuição **normal** N (μ, σ²), a média amostral $\bar{X}$ é exatamente distribuída como N (μ, σ^2/n). Sendo σ **conhecido**, o IC (μ : 1 – α) é dado por:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \text{o qual é construído a partir de} \quad Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad (1)$$

Quando σ é **desconhecido**, como é tipicamente o caso, uma aproximação intuitiva é substituir **σ** por **s** em (1) e considerar a razão:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Essa substituição, embora, não altere consideravelmente a distribuição em amostras grandes, ela causa uma considerável diferença se a amostra for pequena. A notação **t** é requerida porque a variável aleatória no denominador (s) aumenta a variância de **t** para um valor maior do que um (1,0), de modo que a razão não é padronizada.
A distribuição da razão **t**, quando é razoável assumir que a distribuição da população é normal, é conhecida como **distribuição t de Student** com **r = n – 1 graus de liberdade**. A qualificação "n – 1 graus de liberdade" é necessária porque para cada diferente tamanho de amostra (n) ou valor "n – 1", há uma diferente distribuição **t**.

Grau de liberdade (gl) é conceituado como o número de valores independentes de uma estatística. Tomando como exemplo o estimador s^2 de σ^2, foi visto no item 2 que a quantidade (n – 1) é o divisor que aparece na fórmula de s^2. Isto significa que para um tamanho amostral n, $s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$ é baseado em (n – 1) graus de liberdade, ou seja, calculando-se (n – 1) desvios (independentes): $(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_{n-1} - \bar{x})$, o remanescente $(x_n - \bar{x})$ pode ser obtido por diferença, pois $\sum (x_i - \bar{x}) = 0$.
As distribuições **t** são simétricas em torno de zero, $E(t) = 0$, mas têm caudas mais espalhadas, $Var(t) = \frac{r}{r-2} = \frac{n-1}{n-3}$, do que a distribuição *N(0, 1)*. Entretanto, com o aumento de **r**, a distribuição **t** se aproxima da distribuição *N(0, 1)*, pois a *Var(t)* tende a um (1). Assim, quando ***n*** é grande (n ≥ 30), a razão $\frac{\bar{x} - \mu}{s / \sqrt{n}}$, como mencionado anteriormente, é aproximadamente normal padrão. A equivalência entre as distribuições *t* e *N(0, 1)* quando n é grande, pode ser verificada comparando os valores da distribuição *t*, com infinitos (∞) graus de liberdade, com os da normal padrão (Tabelas 3 e 4, respectivamente).

Pode-se concluir da distribuição **t**, que

$$P(-t_{\alpha/2} < \frac{\bar{x} - \mu}{s / \sqrt{n}} < t_{\alpha/2}) = 1 - \alpha, \qquad (2)$$

em que $t_{\alpha/2}$ é obtido na tabela da distribuição *t* com *r = n – 1* graus de liberdade (Tabela 4), a qual fornece valores $t_{\alpha/2}$, tais que $P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1 - \alpha$, para alguns valores de α (ou, como simbolizado na tabela, de *p*) e *r*. Rearranjando os termos dentro dos parênteses da expressão (2), temos

$$P(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}) = 1 - \alpha$$

Portanto, um IC (μ : 1- α) é obtido de $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$. Aqui, o comprimento do intervalo de confiança ($2 \times t_{\alpha/2} \frac{s}{\sqrt{n}}$), tal como no caso em que o tamanho da amostra é grande ($2 \times z_{\alpha/2} \frac{s}{\sqrt{n}}$), é uma variável aleatória, pois envolve o desvio padrão amostral (s). Na situação em que σ é conhecido, ao contrário, todos os intervalos são de mesmo comprimento.

Exemplo 3. Uma amostra de 10 cães sofrendo de uma determinada doença apresentou um tempo de sobrevivência médio de 46,9 meses e o desvio padrão de 43,3 meses. Determinar os limites de confiança de 90% para μ.

Solução: $\bar{x}_a = 46{,}9$ meses s = 43,3 meses

$1 - \alpha = 0{,}90$ n - 1 = 9 $t_{\alpha/2} = 1{,}833$

Limites de confiança para μ: $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 46{,}9 \pm 1{,}833 \frac{43{,}3}{\sqrt{10}} = 21{,}8 \ \ e \ \ 72{,}0 \ meses$

Portanto, IC(μ : 90%) =]21,8; 72,0[

8.4 Intervalo de confiança para o parâmetro binomial p

Fazendo uso do fato que, para ***n*** grande, a distribuição binomial pode ser aproximada com a normal, isto é, que a variável aleatória $Z = \frac{x - np}{\sqrt{np(1-p)}}$ tem distribuição aproximadamente N(0,1), pode-se escrever:

$$P(-z_{\alpha/2} < \frac{x - np}{\sqrt{np(1-p)}} < z_{\alpha/2}) = 1 - \alpha$$

Dividindo-se o numerador e o denominador de Z por n, temos:

$$P(-z_{\alpha/2} < \frac{x/n - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}) = 1 - \alpha \qquad (1)$$

Um intervalo com (1-α)100% de confiança **aproximado** para p é obtido, escrevendo (1) como

$$P(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}) = 1 - \alpha$$

onde $\hat{p}\ (= x/n)$ é a proporção dos elementos da amostra que possuem uma particular característica.

Substituindo *p*, visto que é desconhecido, por seu estimador $\hat{p}$ dentro das raízes, obtêm-se:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Portanto, $$]\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}[$$

é o intervalo de (1 - α)100% de confiança para ***p***. Indica-se por IC (p : 1- α).

O efeito de se utilizar uma estimativa do desvio padrão $\left(\sqrt{\hat{p}\hat{q}/n}\right)$ no IC é desprezível quando *n* é grande (n ≥ 30).

Exemplo 4. Suponha que em *n* = 400 animais são administrados uma droga, obtendo X = 320 sucessos, ou seja, 80% dos animais melhoraram. A partir destes dados, obtenha um IC para p, com 1 - α = 0,90.

Solução: $\hat{p}$ = 320/400 = 0,80 $\hat{q}$ = 0,20

$$\text{IC} = 0{,}80 \pm 1{,}64\sqrt{\frac{0{,}80.0{,}2}{400}} =]0{,}767\,;0{,}833[$$

Portanto, IC(p : 90%) =]0,767 ; 0,833[

8.5 Cálculo do tamanho da amostra

8.5.1 Para estimação de μ

Supondo σ conhecido, o erro da estimação de μ por $\overline{X}$ é $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Fixando um erro máximo de tamanho ***d***, com probabilidade $1-\alpha$, então $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = d$. Resolvendo para n,

$$n = \left[\frac{z_{\alpha/2}\sigma}{d}\right]^2$$

Note que se é σ desconhecido, uma estimativa de σ é necessária para calcular o tamanho da amostra (*n*). Este problema é resolvido por meio de uma amostra preliminar que fornece *s*, que, por sua vez, permite o cálculo de *n*.

Exemplo 5. Um limnologista deseja estimar o conteúdo médio de fosfato por unidade de volume de água de certo lago. Sabe-se de estudos anteriores que *s* = 4. Qual deve ser o tamanho da amostra para que ele tenha 90% de confiança que o erro da estimativa de μ não supere 0,8?

Solução: s = 4 1 - α = 0,90 α/2 = 0,05 $z_{0,05}$ = 1,64 d = 0,8

$$n = \left[\frac{1{,}64.4}{0{,}8}\right]^2 = 67{,}24 \approx 68$$

8.5.2 Para estimação de p

Neste caso, $d = z_{\alpha/2}\sqrt{\frac{p\,q}{n}}$. Assim, $n = p\,q\left[\frac{z_{\alpha/2}}{d}\right]^2$.

Esta solução não é usada, porque ela envolve o parâmetro p, que é desconhecido. Os valores de p variam de 0 a 1, de modo que p (1 - p) aumenta de 0 até 1/4 (valor máximo), decrescendo, a partir daí, até 0. O valor máximo de pq é 1/4, quando p = q = 1/2, de modo que a solução n deve satisfazer

$$n \leq \frac{1}{4}\left[\frac{z_{\alpha/2}}{d}\right]^2$$

Sem qualquer conhecimento prévio do valor aproximado de p, a escolha do n máximo proporciona a proteção desejada. Se for conhecido que o valor de p está próximo de um valor p*, então n pode ser determinado de

$$n = p^*(1-p^*)\left[\frac{z_{\alpha/2}}{d}\right]^2$$

Exemplo 6. A inspeção de saúde pública foi designada para estimar a proporção p de uma população bovina tendo certa anomalia infecciosa. Quantos animais devem ser examinados (tamanho da amostra) para que se tenha 98% de confiança de que o erro da estimativa não seja superior a 0,05, quando (a) não há conhecimento a cerca do valor de p? e (b) sabe-se que p é aproximadamente 0,3?

Solução:

$d = 0{,}05 \quad 1 - \alpha = 0{,}98 \quad \alpha/2 = 0{,}01 \quad z_{0,01} = 2{,}33$

(a) $$n = p(1-p)\left[\frac{z_{\alpha/2}}{d}\right]^2 = \frac{1}{4}\left[\frac{2{,}33}{0{,}05}\right]^2 = 543$$ para p = q = 1/2 (n máximo)

(b) $$n = 0{,}3.0{,}7\left[\frac{2{,}33}{0{,}05}\right]^2 = 456$$

2.3. Para estimação de μ em populações finitas (amostra "sem reposição")

Supondo uma população com N elementos,

$$d = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad \Rightarrow \qquad \sqrt{n} = \frac{z_{\alpha/2}\,\sigma\sqrt{\frac{N-n}{N-1}}}{d}$$

$$n = \frac{z^2_{\alpha/2}\sigma^2 \frac{N-n}{N-1}}{d^2} \qquad \Rightarrow \qquad n = z^2_{\alpha/2}\sigma^2 \frac{N-n}{N-1} \cdot \frac{1}{d^2}$$

$$n(N-1)d^2 + z^2_{\alpha/2}\sigma^2 n = z^2_{\frac{\alpha}{2}}\sigma^2 N \qquad \Rightarrow \qquad n[(N-1)d^2 + z^2_{\alpha/2}\sigma^2] = z^2_{\alpha/2}\sigma^2 N$$

Portanto, $$n = \frac{z^2_{\alpha/2}\sigma^2 N}{(N-1)d^2 + z^2_{\alpha/2}\sigma^2} \qquad (1)$$

Por exemplo, nas condições do Exemplo 5 e considerando N =1000:

$$n = \frac{z^2_{\alpha/2}\sigma^2 N}{(N-1)d^2 + z^2_{\alpha/2}\sigma^2} = \frac{1{,}64^2 \cdot 16 \cdot 1000}{999 \cdot 0{,}8^2 + 1{,}64^2 \cdot 16} \approx 63$$

Note que em (1) quando d for pequeno, por exemplo, d = 0,03, o termo $(N - 1)d^2$ também será pequeno, logo o tamanho da amostra (n) será aproximadamente igual ao da população (N).

2.4. Para estimação de p em populações finitas (amostra "sem reposição")

Supondo uma população com N elementos,

$$d = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Para p = q = 0,5

$$d = z_{\alpha/2}\sqrt{\frac{0{,}25}{n}}\sqrt{\frac{N-n}{N-1}} \qquad \Rightarrow \qquad d^2 = z^2_{\alpha/2}\frac{0{,}25}{n}\frac{N-n}{N-1}$$

$$d^2 = 0{,}25 z^2_{\alpha/2}\frac{N-n}{n(N-1)} \qquad \Rightarrow \qquad \frac{d^2}{0{,}25 z^2_{\alpha/2}} = \frac{N-n}{n(N-1)}$$

$$\frac{d^2}{0{,}25 z^2_{\alpha/2}} n(N-1) = N-n \qquad \Rightarrow \qquad [\frac{d^2}{0{,}25 z^2_{\alpha/2}} n(N-1)] + n = N$$

$n\{[\frac{d^2}{0{,}25 z^2_{\alpha/2}}(N-1)] + 1\} = N$. Portanto,

$$n = \frac{N}{[\frac{d^2}{0{,}25 z^2_{\alpha/2}}(N-1)] + 1} \qquad (2)$$

Por exemplo, nas condições do Exemplo 6 e considerando N = 1000:

$$n = \frac{1000}{(\frac{0{,}05^2}{0{,}25.2{,}33^2} \cdot 999) + 1} = \frac{1000}{(0{,}00184.999) + 1} = \frac{1000}{2{,}84} \approx 352$$

Note que em (2) quando d for pequeno, por exemplo, d = 0,003 (0,3%), o termo $[\frac{d^2}{0{,}25 z_{\alpha/2}^2}(N-1)] + 1$ também será pequeno, logo o tamanho da amostra (n) será aproximadamente igual ao da população (N).

8.5.3 Para estimação de p usando probabilidades binomiais b(x : n, p)

Quando a ocorrência de certa característica em uma população é pouco frequente, podemos calcular o tamanho da amostra (n) para a estimação de p, considerando uma probabilidade para que tenhamos pelo menos um (1) sucesso (S) na amostra, que seja maior ou igual a β (%). Essa probabilidade binomial, em termos matemáticos, pode ser representada por:

P (pelo menos 1 S) = 1 – P (nenhum S) = 1 – P (X = 0) ≥ β

P (pelo menos 1 S) = 1 – P (nenhum S) $= 1 - \binom{n}{0} p^0 q^n \geq \beta$

Logo, $1 - q^n \geq \beta \Rightarrow -q^n \geq \beta - 1 \Rightarrow q^n \leq 1 - \beta$ (1)

Aplicando-se logaritmo em ambos lados de (1), obtêm-se: $\ln q^n \leq \ln(1-\beta)$ (2)

Resolvendo (2) para n, $n \geq \frac{\ln(1-\beta)}{\ln q}$

Por exemplo, se P (S) = p = 0,1 e β = 90 %

$$n \geq \frac{\ln 0{,}10}{\ln 0{,}90} \Rightarrow n \geq \frac{-2{,}302}{-0{,}105} \Rightarrow n \geq 22$$

e se p = 0,01, $n \geq \frac{\ln 0{,}10}{\ln 0{,}99} \Rightarrow n \geq \frac{-2{,}302}{-0{,}010} \Rightarrow n \geq 230$

Exemplo 7. Uma doença em bovinos torna-se grave, quando ocorre acima de um certo limite. Qual deve ser o tamanho da amostra (n) para detectar a presença dessa doença com 95 % (β) de segurança, quando a mesma está presente em 10 % (p) dos animais?

Solução:

$$n = \frac{\ln 0{,}05}{\ln 0{,}90} = \frac{-2{,}996}{-0{,}105} \cong 28$$

9 TESTES DE HIPÓTESES

Aqui estudaremos outro aspecto da inferência estatística: o teste de hipóteses, cujo o objetivo é decidir se uma afirmação, em geral, sobre **parâmetros** de uma ou mais populações é, ou não, apoiado pela evidência obtida de dados amostrais. Tal afirmação é o que se chama **Hipótese Estatística** e a regra usada para decidir se ela é verdadeira ou não, é o **Teste de Hipóteses**. Iremos ilustrá-lo por meio de um exemplo.

Exemplo 1. Uma suinocultura usa uma ração A que propicia, da desmama até a idade de abate, um ganho em peso de 500 g/dia/suíno (σ = 25 g). O fabricante de uma ração B afirma que nas mesmas condições, sua ração propicia um ganho de 510 g/dia (σ = 25 g). É evidente que em termos financeiros, se for verídica a afirmação do fabricante da ração do tipo B, esta deve ser usada em substituição à do tipo A.

Se o criador tem de decidir com base em uma amostra, se o ganho em peso dos suínos dando a nova ração é 510 g/dia, o problema pode ser expresso na linguagem de teste estatístico de hipóteses.

9.1 Hipóteses estatísticas

Em experimentos comparativos, nos quais um novo produto ou nova técnica é comparado com o padrão, para determinar se sua superioridade pode ser corroborada pela evidência experimental, é necessário formular a:

Hipótese nula (H_0), cujo termo é aplicado para a hipótese a ser testada, e a

Hipótese alternativa (H_1)

A hipótese nula (H_0) é a hipótese de **igualdade** entre o novo e o produto padrão, ou seja, a designação "hipótese nula" decorre da suposição que a diferença entre eles é nula ou zero.

A análise de cada situação indicará qual deve ser considerada a hipótese nula e qual a hipótese alternativa. Uma especificação de H_0 e H_1 no exemplo seria:

$$\begin{cases} H_0 : \mu = 500 \text{ g/dia (a ração B não é melhor)} \\ H_1 : \mu = 510 \text{ g/dia (a ração B é melhor)} \end{cases}$$

ou

$H_0 : \mu = \mu_0$
$H_1 : \mu = \mu_1$

onde: $\mu_1 > \mu_0$ e $\sigma = 25$

Se uma hipótese estatística especifica o valor do parâmetro, ela é referida como hipótese **simples**; se não, é referida como **composta**. Assim, no exemplo, a hipótese alternativa μ = 510 é **simples**. Seria **composta**, por exemplo, se $\mu > 500$, visto que não fixa um valor específico para o parâmetro μ. Em H_0, o valor do parâmetro tem de ser especificado.

A hipótese preferencial é H_0 e é sustentada como **verdadeira**, a menos que os dados se coloquem firmemente contra ela. Em tal caso, H_0 seria rejeitada a favor de H_1. Rejeitar erradamente H_0 é visto como um erro mais grave do que não rejeitar H_0 quando H_1 é verdadeira.

9.2 Erros tipos I e II

O problema proposto consiste em verificar se com a utilização da nova ração, a média de ganho em peso seria estatisticamente maior que 500 g e caso isto se verifique, a suinocultura passaria a utilizá-la. Caso contrário, continuaria com a ração do tipo A, que já foi testada (conhecida *a priori*).

Para a tomada de decisão, deve-se extrair uma amostra aleatória (por exemplo, n = 50) de suínos, fornecendo à mesma, da desmama até a idade de abate, a ração B, e após o término da prova, calcula-se a média amostral ($\bar{x}_a$) do ganho diário em peso no período, que é, no caso, a **estatística teste**. A estatística teste é o valor amostral da estatística utilizada para testar um parâmetro no teste de hipóteses.

Parece razoável estabelecer que se $\bar{x}_a$ estiver próxima de 500 g, não se deve rejeitar H_0, e a conclusão é que a ração do tipo B é estatisticamente igual a do tipo A. Por outro lado, se $\bar{x}_a$ estiver próxima ou for superior à 510 g, a tomada

de decisão é que a ração do tipo B é superior à do tipo A (rejeitar H_0) e que a suinocultura passe a utilizá-la. A média amostral ($\bar{x}_a$) é, no entanto, uma variável aleatória que pode assumir qualquer valor entre 500 e 510 g. Assim, deve-se estabelecer um critério de decisão para aceitar ou rejeitar H_0. Isto é feito determinando um valor **k** (ponto) entre 500 e 510 g, chamado **valor crítico** $(\bar{x}_c)$, e adotando a seguinte regra de decisão:

"Se a média amostral ($\bar{x}_a$) estiver à direita de k, rejeita-se H_0, caso contrário não se rejeita"

Graficamente tem-se a seguinte situação:

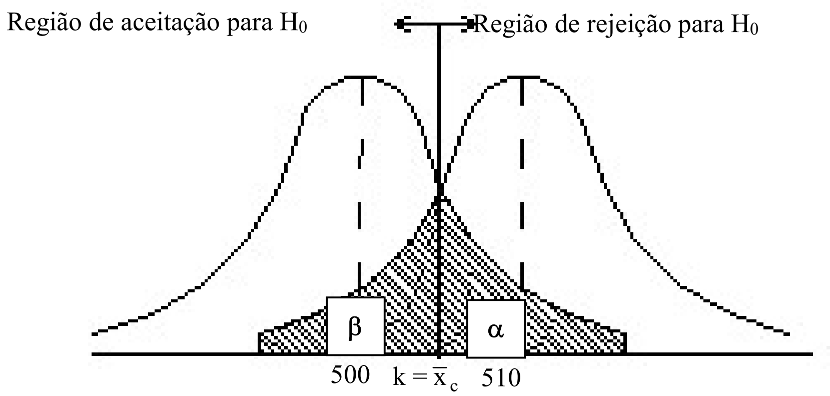


Figura 1. Região de rejeição de H_0 para o teste $\mu = \mu_0$ vs. $\mu = \mu_1$

Um teste de hipóteses é completamente especificado pela estatística teste e região de rejeição. A região de rejeição ou região crítica (RC) é o conjunto de valores da estatística teste para os quais H_0 é rejeitada.

O procedimento do teste, então, divide os possíveis valores da estatística teste em dois subconjuntos: uma região de aceitação e uma de rejeição para H_0, o que pode levar a dois tipos de erros. Por exemplo, se o verdadeiro valor do parâmetro μ é 500 g e incorretamente concluímos que μ = 510 g, cometeremos um erro referido como **erro tipo I**. Por outro lado, se o verdadeiro valor de μ é 510 g e incorretamente concluímos que μ = 500 g, cometeremos uma segunda espécie de erro, referido como **erro tipo II**.

O quadro abaixo resume a natureza dos erros envolvidos no processo de decisão, por meio dos testes de significância:

Conclusão do teste	Situação específica na população	
	H_0 verdadeira	H_0 falsa
Não rejeitar H_0	Decisão correta	Erro tipo II (perdas potenciais para o criador)
Rejeitar H_0	Erro tipo I (perdas reais para o criador)	Decisão correta

Denota-se por:

α = P (erro tipo I) = P (rejeitar H_0/H_0 é verdadeira)

β = P (erro tipo II) = P (não rejeitar H_0/H_0 é falsa)

Assim, o tamanho da região crítica é exatamente a probabilidade α de cometer o erro tipo I . Essa probabilidade é também chamada de **nível de significância** do teste. O nível de significância do teste (α) é, portanto, a probabilidade com que desejamos correr o risco de cometer o erro tipo I, ou seja, em α% dos casos de rejeição de H_0, estaremos tomando decisão errada.

Escolhendo um valor para $\bar{x}_c$, pode-se determinar as probabilidades α e β de cometer cada tipo de erro. Mas, o procedimento que se usa na prática para construir a regra de decisão é fixar α, a probabilidade do erro tipo I (rejeitar H_0 quando ela for verdadeira). O valor é arbitrário e o resultado da amostra é tanto mais significante para rejeitar H_0 quanto menor for esse nível. Geralmente, o valor é fixado em 5%, 1% ou 0,1%.

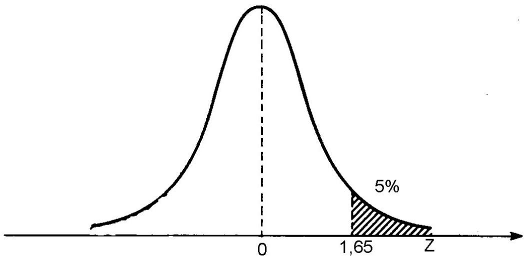
Por exemplo, fixemos α em 5%, ou seja, P(erro I) = P($\bar{X} \geq \bar{x}_c / H_0$ é verdadeira) = 5%, e vejamos qual a regra de decisão correspondente.

Quando H_0 é verdadeira (μ = 500 g), sabe-se do Teorema Limite Central, que $\bar{X}$, a média de amostras de tamanho 50, terá distribuição aproximadamente

$N[\mu(=500); \dfrac{\sigma^2(=625\,g^2)}{n(=50)}]$ ou seja, $N(500\,g; 12{,}5\,g^2)$. Assim,

$$P(erro\,I) = P[\bar{X} \geq \bar{x}_c / \bar{X} : N(500\,g; 12{,}5\,g^2)] = 5\%$$

$$P[Z \geq \frac{\bar{x}_c - \mu_0}{\sigma/\sqrt{n}}] = P[Z \geq \frac{\bar{x}_c - 500}{3{,}5}] = 5\% \qquad \Rightarrow \qquad \frac{\bar{x}_c - 500}{3{,}5} = 1{,}65$$



ou seja, $\bar{x}_c = k = (3{,}5.1{,}65) + 500 = 505{,}78g$

Então, RC = {$\overline{X} \in$ R/ $\overline{X} \geq$ 505,78 g} e a regra de decisão é: "se $\bar{x}_a \in$ RC, rejeita-se H_0 e a conclusão é que a ração B é superior à A; se $\notin$, não se rejeita H_0, e a conclusão é que as rações são estatisticamente iguais".

Convém observar que a RC é sempre construída usando os valores hipotetizados por H_0 ou seja, sob a hipótese H_0 ser verdadeira.

Com essa regra de decisão:

β = P(erro II) = P[$\overline{X}$ < 505,78/ $\overline{X}$: N(510 g, 12,3 g^2)]

β = P [Z < $\dfrac{505{,}78 - 510}{3{,}5}$] = P[Z < -1,21] = 11,31 %

Há uma relação inversa entre α e β, ou seja, se a probabilidade de um tipo de erro é reduzida, aquela do outro tipo é aumentada (Verifique na Figura 1). No caso da escolha de um valor para $\bar{x}_c$, por exemplo, 505 kg (o ponto médio entre 500 e 510 kg), pode-se reduzir as probabilidades de ambos os tipos de erros, aumentando o tamanho da amostra (**n**). Este resultado também pode ser facilmente verificado a partir da Figura 1, considerando que, da transformação para a normal reduzida, $z_c = \dfrac{\bar{x}_c - \mu}{\sigma/\sqrt{n}}$.

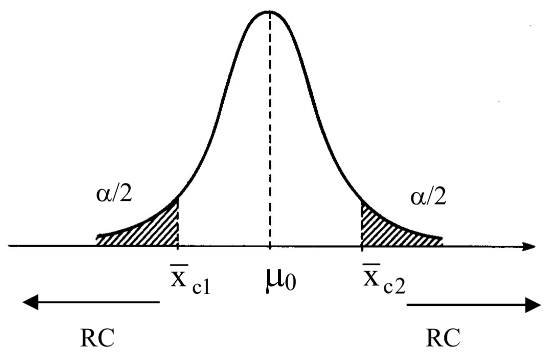
A probabilidade com que o teste de significância, com α fixado, rejeita H_0, quando o particular valor alternativo do parâmetro é verdadeiro, é chamada **poder do teste**. O poder do teste é um menos a probabilidade do erro tipo II ou seja, (1 - β). No exemplo, o poder do teste é: **1 - β** = 1 – 0,1131 = 0,8869 (88,7%).

Frequentemente, no entanto, não são especificados valores **fixos** para o parâmetro em H_1. Então, sua caracterização dependerá do grau de conhecimento que se tem do problema. A alternativa mais geral é:

H_1: $\mu \neq \mu_0$ (teste bilateral)

Neste caso, a regra de decisão deverá indicar dois pontos $\bar{x}_{c1}$ e $\bar{x}_{c2}$, tais que, H_1 será sustentada se a média da amostra for muito grande ou muito pequena. Então, a estrutura apropriada da região de rejeição ou crítica (RC) é:

"rejeita-se H_0 se $\overline{X} \leq \bar{x}_{c1}$ *ou* $\overline{X} \geq \bar{x}_{c2}$"



Com esta regra de decisão, não podemos encontrar β, consequentemente, não podemos controlar o erro tipo II, pois o valor do parâmetro sob a hipótese alternativa não é especificado.
Voltando ao problema proposto, e testando

H_0: μ = 500 g vs. H_1: μ ≠ 500 g

tem-se, fixando α = 5%,

P(erro I) = P[$\overline{X} \leq \overline{x}_{c1}$ ou $\overline{X} \geq \overline{x}_{c2}$ / $\overline{X}$: N (500 g, 12,3 g²)] = 5%

= P[Z ≤ -1,96 ou Z ≥ 1,96) = 5%

$$-1,96 = \frac{\overline{x}_{c1} - 500}{3,5} \quad \therefore \quad \overline{x}_{c1} = 493,1\,\text{g}$$

$$1,96 = \frac{\overline{x}_{c2} - 500}{3,5} \quad \therefore \quad \overline{x}_{c2} = 506,9\,\text{g}$$

Assim,

RC = {$\overline{X}$ ∈ R/ $\overline{X}$ ≤ 493,1g ou $\overline{X}$ ≥ 506,9 g}

A extensão para testes unilaterais das formas:

H_1: μ > μ_0 (teste unilateral à direita) e

H_1: μ < μ_0 (teste unilateral à esquerda), é imediata.

Exemplo 2. No caso da suinocultura, considerando a amostra de 50 leitões (n = 50), aos quais foi fornecida a nova ração (B), deve-se ou não adotar essa ração, admitindo-se como resultado um ganho em peso médio diário de 504 g ($\overline{x}_a = 504g$), fixando α = 5%?

Solução:

H_0: μ = 500 g
H_1: μ = 510 g
$\overline{x}_a = 504g$ n = 50 α = 0,05 σ = 25 g

$$z_c = \frac{\overline{x}_c - \mu_0}{\sigma / \sqrt{n}} \quad \Rightarrow \quad 1,65 = \frac{\overline{x}_c - 500}{25/\sqrt{50}} \quad \therefore \quad \overline{X}_c = 505,78\text{ g}$$

RC = {$\overline{X} \geq 505{,}78$ g}

Conclusão:

Como $\overline{x}_a \notin$ RC, não se rejeita H_0 ao nível de significância de 5%, ou seja, a ração B não é melhor do que a A. Portanto, a suinocultura não deve adotá-la.

Equivalentemente, os testes descritos podem ser baseados na estatística:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$, obtendo-se as regiões críticas na distribuição N (0,1).

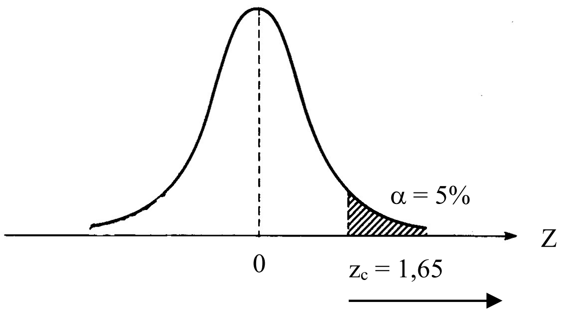
Esta expressão corresponde à seguinte fórmula geral:

$$Estatística\, teste = \frac{estimativa\, do\, parâmetro - valor\, do\, parâmetro\, hipotetizado\, por\, H_0}{erro\, padrão\, da\, estimativa\, do\, parâmetro},$$

que será aplicada daqui em diante em testes de hipóteses.

Assim procedendo na resolução do Exemplo 2, o valor observado da estatística teste (Z_{obs}) é dado por:

$$z_{obs} = \frac{\overline{x}_a - \mu_0}{\sigma / \sqrt{n}} = \frac{504 - 500}{25 / \sqrt{50}} = 1{,}14$$



RC = {$Z \geq 1{,}65$}

Como $z_{obs} < z_c$, não se rejeita H_0 ao nível de 5%.

9.3 Passos para a construção de um teste de hipóteses

Nos itens anteriores foram introduzidos os conceitos básicos e as terminologias que são aplicados em testes de hipóteses. Um sumário dos principais passos que podem ser usados sistematicamente para qualquer teste de hipóteses é apresentado aqui, ou seja:

(a) Fixe a hipótese H_0 a ser testada e a alternativa H_1;

(b) Use a teoria estatística e as informações disponíveis para decidir qual estatística (estimador) será usada para testar a hipótese H_0, obtendo-se suas propriedades (distribuição, estimativa, erro padrão);
(c) Fixe a probabilidade α de cometer o erro tipo I e use este valor para construir a RC (região crítica). Lembre-se que a RC é construída para a estatística definida no passo (a), usando os valores hipotetizados por H_0;
(d) Use as informações da amostra para calcular o valor da estatística do teste; e
(e) Se o valor da estatística calculado com os dados da amostra não pertencer à RC, não rejeite H_0; caso contrário, rejeite H_0.

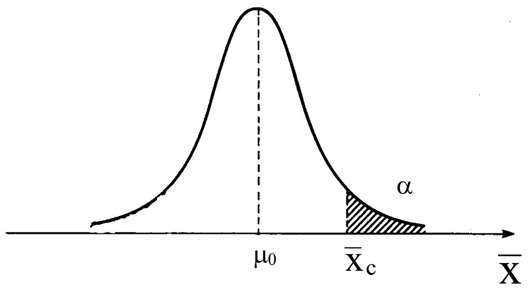
9.4 Teste sobre a média de uma população com variância conhecida

Descreveremos agora, de modo sucinto, os passos básicos definidos na seção anterior, para testar a hipótese de que a média de uma população μ é igual a um número fixado μ_0, supondo que a população tem distribuição normal, cuja variância (σ^2), embora seja uma condição irreal, é conhecida.

Hipótese simples vs. alternativa simples

(a) Teste unilateral à direita

$H_0 : \mu = \mu_0$
$H_1 : \mu = \mu_1 \quad (\mu_1 > \mu_0)$



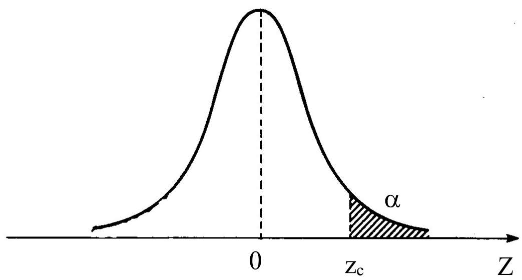
Com α fixado,

RC = $\{\overline{X} \in R / \overline{X} \geq \bar{x}_c\}$, onde: $\bar{x}_c$ é obtido a partir de $z_c = \dfrac{\bar{x}_c - \mu_0}{\sigma / \sqrt{n}}$,

sendo z_c: N(0,1), tal que $P(Z \geq z_c) = \alpha$

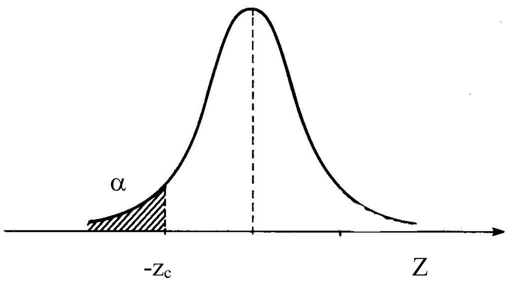
Equivalentemente,

RC = $\{Z \geq z_c\}$, onde: $Z = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$



(b) Teste unilateral à esquerda

$H_0 : \mu = \mu_0$
$H_1 : \mu = \mu_1 \qquad (\mu_1 < \mu_0)$



$RC = \{Z \leq - z_c\}$

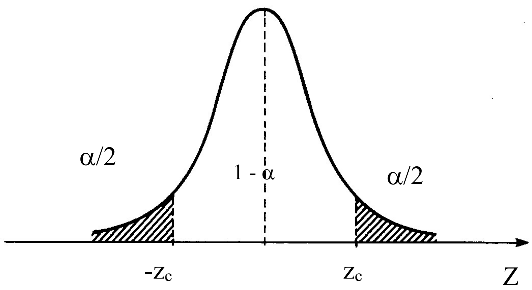
4.2. Hipótese simples vs. alternativa composta

(i)	$H_0 : \mu = \mu_0$	
	$H_1 : \mu > \mu_0$	RC idêntica à de (a)

(ii)	$H_0 : \mu = \mu_0$	
	$H_1 : \mu < \mu_0$	RC idêntica à de (b)

(iii)	$H_0 : \mu = \mu_0$
	$H_1 : \mu \neq \mu_0$

Teste bilateral da forma:



$RC = \{Z \geq z_c \text{ ou } Z \leq - z_c\}$

Exemplo 3. Usando os dados do Exemplo 1, testar a hipótese de μ = 500 g contra a hipótese alternativa $\mu \neq$ 500 g, ao nível de significância de 5%.
Solução:
H_0: μ = 500 g $\qquad \bar{x}_a = 504\,g \qquad \alpha = 5\%$
H_1: $\mu \neq$ 500 g

RC = {Z ≥ 1,96 ou Z ≤ - 1,96} $$z_{obs} = \frac{\bar{x}_a - \mu_0}{\sigma/\sqrt{n}} = \frac{504 - 500}{25/\sqrt{50}} = 1{,}14$$

Conclusão:
Como $z_{obs} \notin$ RC, não se rejeita H_0 ao nível de 5%, ou seja, a ração B não é estatisticamente melhor do que a A.

9.5 Probabilidade de significância (valor-p)

Existem duas opções para expressar a conclusão final de um teste de hipóteses:
- Comparar, como descrito anteriormente, o valor da estatística teste com o valor obtido a partir da distribuição teórica, específica para o teste, para um valor pré-fixado do nível de significância (α);
- Quantificar a chance do que foi observado ou resultados mais extremos, sob a hipótese nula (H_0) ser verdadeira. Essa opção baseia-se na probabilidade de ocorrência de valores iguais ou superiores ao assumido pela estatística teste, dado que a hipótese H_0 é verdadeira. Este número é chamado de probabilidade de significância ou valor-p e frequentemente é indicado apenas por *p*.

Obs. *Valor-p* e nível de significância (α) não são sinônimos. O *valor-p* é sempre obtido de uma amostra, enquanto o nível de significância é geralmente fixado antes da coleta dos dados.

Definição: *valor-p*, também denotado como nível descritivo do teste, é o nome que se dá à probabilidade de se observar um resultado tão ou mais extremo que o da amostra, supondo que a hipótese nula seja verdadeira. No caso de um teste de hipóteses no qual o valor da estatística teste é Z_{obs}, o *valor-p* é dado por:
$p = P(Z \geq Z_{obs} \mid H_0)$.

Em outras palavras, o valor-p corresponde ao menor nível de significância que pode ser assumido para rejeitar a hipótese nula. ***Dizemos então que há significância estatística quando o valor-p é menor que o nível de significância adotado*** (α).

Para exemplificar a definição de valor-p, consideremos primeiro o caso de um teste de hipóteses monocaudal para a média. Vide Exemplo 2, onde $\alpha = 0{,}05$ e $Z_{obs} = 1{,}14$. Assim,

$p = P(Z \geq Z_{obs}) = P(Z \geq 1{,}14) = 0{,}12714$

Portanto, podemos concluir que, para qualquer nível de significância maior que 0,12714, temos evidências para rejeitar a hipótese nula. Observe que o valor-p é maior que o nível de significância proposto $(p > \alpha)$, assim, como concluído, não rejeitamos a hipótese nula (H_0: μ = 500 g). Além disso, quanto maior (ou menor) for o *valor-p*, mais "próximo" (ou "distante") estamos da hipótese nula (H_0). Do que se deduz que o *valor-p* tem mais informações sobre a evidência contra hipótese H_0 e deste modo o experimentador tem mais informações para decidir sobre ela, com o nível de significância apropriado. Ao contrário, se o *valor-p* for menor que o nível de significância proposto $(p < \alpha)$, rejeita-se H_0.

Considerando agora o teste para a média como bicaudal (vide Exemplo 3), segue que o *valor-p* é dado por:

p = $P(Z \geq Z_{obs}) + P(Z \leq -Z_{obs}) = P(Z \geq 1{,}14) + P(Z \leq -1{,}14) = 0{,}2542$

donde podemos concluir que, para qualquer nível de significância menor que 0,2542, temos evidências, como no caso do exemplo, para não rejeitar a hipótese nula.

Em geral, os resultados podem ser interpretados como:
Valor-p próximo de 0 - Um indicador de que a hipótese nula é falsa.
Valor-p próximo de 1 - Não há evidência suficiente para rejeitar a hipótese nula.

Normalmente considera-se um *valor-p* de 0,05 como o patamar para avaliar a hipótese nula (H_0). Se o valor-p for inferior a 0,05 podemos rejeitar H_0. Em caso contrário, não temos evidência que nos permita rejeitá-la (o que não significa automaticamente que seja verdadeira). Em situações de maior exigência é usado um valor-p inferior a 0,05. Na maioria dos softwares, a significância estatística é expressa pelo nível descritivo (valor-p).

9.6 Teste para proporção

Considere uma população e uma hipótese sobre uma proporção **p** dessa população:

$H_0 : p = p_0$

O problema fornece informações sobre H_1, que pode ser:

(a) $H_1 : p = p_1$ $\quad p_1 > p_0$ (teste monocaudal à direita)
(b) $H_1 : p = p_1$ $\quad p_1 < p_0$ (teste monocaudal à esquerda)
(c) $H_1 : p > p_0$ (teste monocaudal à direita)
(d) $H_1 : p < p_0$ (teste monocaudal à esquerda)
(e) $H_1 : p \neq p_0$ (teste bicaudal)

Quando **n** (tamanho da amostra) é grande,

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

onde: $\hat{p}$ é a proporção da amostra

Sob H_0 verdadeira,

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$$

e para todas as formas de H_1

$$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N\ (0,1)$$

As regiões críticas são idênticas às mostradas em (3) e os valores de z_c, fixando-se α, são obtidos na distribuição N (0,1).

Exemplo 4. Um laboratório de vacinas contra febre aftosa reivindicou que ela imuniza 90% dos animais. Em uma amostra de 200 animais, nos quais foram aplicados a vacina, 160 foram imunizados. Verificar se a declaração do fabricante é verdadeira ao nível de 5%.

Solução:

$H_0 : p = 0{,}90\ (p_0)$

$H_1 : p < 0{,}90$

$n = 200 \qquad \hat{p} = \frac{160}{200} = 0{,}80 \qquad \alpha = 0{,}05$

$$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0{,}80 - 0{,}90}{\sqrt{(0{,}90.0{,}10)/200}} = -\ 4{,}72$$

$RC = \{Z \leq -1{,}65\}$

Decisão:

Como $z_{obs} < z_c$, rejeita-se H_0 ao nível de 5%, ou seja, a proporção de imunização é menor do que 90%.

Conclusão:

A declaração do laboratório é falsa ao nível de 5%.

9.7 Teste para a média de uma população N(μ, σ^2), σ^2 desconhecido

Hipóteses:

H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$ [ou $\mu > \mu_0$ ou $\mu < \mu_0$], onde μ_0 é um valor conhecido.

Estatística teste: Neste caso, a exemplo do que foi feito na construção de intervalos de confiança, a estatística a ser usada para testar a hipótese H_0 é:

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

que tem distribuição *t de Student* com *n –1* graus de liberdade (t_{n-1}).

Região crítica: Fixado α , a região crítica (RC) é:

$$RC: t_{n-1} < -t_{\alpha/2,\, n-1} \;\; ou \;\; t_{n-1} > t_{\alpha/2,\, n-1}$$

ou $RC: |t_{n-1}| > t_{\alpha/2,\, n-1}$.

Os valores de $t_{\alpha/2,\, n-1}$ podem ser obtidos na Tabela 4, apresentada no capítulo anterior.

Resultado da amostra: Colhida uma amostra aleatória de tamanho n, calculada sua média $(\bar{x}_a)$ e desvio padrão (s_a), calcula-se:

$$t_{obs} = \frac{\bar{x}_a - \mu_0}{s_a / \sqrt{n}}$$

Análise do resultado: Se $t_{obs} \in$ RC, **rejeita-se** H_0; caso contrário, **não se rejeita**

Esse teste é chamado ***teste t de Student*** ou, simplesmente, ***teste t***.

Se n for grande ($n \geq 30$), $\bar{x}$, como já visto, pode ser tratada como uma variável aproximadamente normal $N(\mu, \sigma^2/n)$, em virtude da aplicação do teorema limite central. Além disso, σ pode ser substituído por *s* sem afetar consideravelmente a distribuição. Assim, um teste aproximado de H_0: $\mu = \mu_0$ pode ser executado usando-se a estatística Z, consultando a tabela normal para a região de rejeição.

Exemplo 5. As especificações de uma dada droga veterinária exigem 23,2 g de álcool etílico. Uma amostra de 10 análises do produto apresentou um teor médio de álcool de 23,5g com desvio padrão de 0,24g. Pode-se concluir ao nível de significância de 1% que o produto satisfaz as condições exigidas ($\mu = 23{,}2$g)?
Solução:

H_0: $\mu = 23{,}2$ g
H_1: $\mu \neq 23{,}2$ g

$\alpha = 0{,}01$ $\quad \bar{x}_a = 23{,}5\,g$ $\quad s_a = 0{,}24$ $\quad n = 10$

Consultando a Tabela 4, $t_{c(0,01;\,9)} = 3{,}25$, de modo que
RC = {t < -3,25 ou t > 3,25}

$$t_{obs} = \frac{\bar{x}_a - \mu_0}{s_a / \sqrt{n}} = \frac{23{,}5\,g - 23{,}2\,g}{0{,}24 / \sqrt{10}} = 3{,}95$$

Conclusão: como $t_{obs} \in$ RC, rejeita-se H_0 ao nível de 1%, ou seja, o teste indica que o produto não satisfaz as condições exigidas.

10 COMPARAÇÕES DE PARÂMETROS DE DUAS POPULAÇÕES

10.1 Comparação das variâncias de duas populações normais

Suponha duas amostras aleatórias independentes de tamanhos n_1 e n_2 ou seja, $X_1, X_2, ..., X_{n_1}$ e $Y_1, Y_2, ..., Y_{n_2}$, respectivamente, de uma população com distribuição $N(\mu_1, \sigma_1^2)$ e de uma população com distribuição $N(\mu_2, \sigma_2^2)$.

Hipóteses:

$H_0: \sigma_1^2 = \sigma_2^2 \qquad (\text{ ou } \sigma_1^2 / \sigma_2^2 = 1)$

$H_1: \sigma_1^2 \neq \sigma_2^2 \qquad (\text{ ou } \sigma_1^2 / \sigma_2^2 \neq 1)$

Estatística do teste:

Sendo s_1^2 e s_2^2 as variâncias, respectivamente, das amostras n_1 e n_2, o quociente

$$\frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

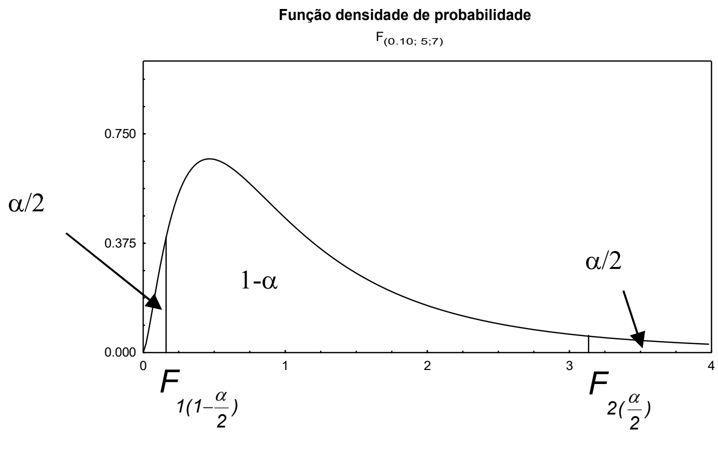
segue a distribuição de F (Snedecor) com n_1-1 e n_2-1 graus de liberdade (gl) [F(n_1-1, n_2-1)].

Sob a suposição de H_0 ser verdadeira, isto é, $\sigma_1^2 = \sigma_2^2$, tem-se que

$$F = \frac{s_1^2}{s_2^2} : F(n_1 - 1,\ n_2 - 1)$$

Construção da região crítica:

Fixado α, os pontos críticos serão F_1 e F_2 da distribuição F, tais que :



Se α = 10%, pode-se, utilizando a Tabela 5, encontrar diretamente $F_{2(5\%)}$. Para encontrar $F_{1(95\%)}$ utiliza-se a propriedade:

$$F_{(1-\alpha;\, n_1-1,\, n_2-1)} = \frac{1}{F_{(\alpha;\, n_2-1,\, n_1-1)}} = F_{(0,95;\, n_1-1,\, n_2-1)} = \frac{1}{F_{(0,05;\, n_2-1,\, n_1-1)}}$$

Por exemplo, se n_1-1 = 5 e n_2 -1 = 7,

$$F_{2(0,05;5,7)} = 3{,}97$$

$$F_{1(0,95;\,5,7)} = \frac{1}{F_{(0,05;7,5)}} = \frac{1}{4{,}88} = 0{,}205$$

Assim, RC = { 0 < F < 0,205 ou F > 3,97 }

Entretanto, o procedimento que se usa na prática é calcular F utilizando sempre a maior variância no numerador ($s_1^2 > s_2^2$), portanto F > 1, e considerar o ponto crítico $F_{2(\alpha/2;n_1-1,n_2-1)}$.

Amostra: Colhidas amostras aleatórias n_1 e n_2, calcula-se s_1^2 e s_2^2 ($s_1^2 > s_2^2$), então

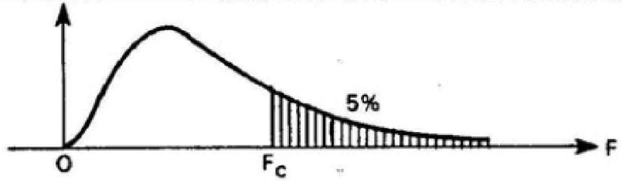
$$F_{obs} = \frac{s_1^2}{s_2^2} : F(n_1-1, n_2-1)$$

Conclusão: Se $F_{obs} \in$ RC, **rejeita-se** H_0, caso contrário, **não se rejeita**.

TÁBUA VI

Distribuição de Fisher-Snedecor-1

Valores críticos de F tais que $P(F > F_c) = 0{,}05$



Graus de liberdade do denominador de F : n_2	1	2	3	4	5	6	7	8	9	10	12	14	15	16	18	20	24	30	40	60	120	∞	Graus de liberdade do denominador de F : n_2
	GRAUS DE LIBERDADE DO NUMERADOR DE F : n_1																						
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,4	245,9	246,5	247,3	248,0	249,1	250,1	251,1	252,2	253,3	254,3	1
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,42	19,43	19,43	19,44	19,45	19,45	19,46	19,47	19,48	19,49	19,50	2
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,72	8,70	8,69	8,67	8,66	8,64	8,62	8,59	8,57	8,55	8,53	3
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,87	5,86	5,84	5,82	5,80	5,77	5,75	5,72	5,69	5,66	5,63	4
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,64	4,62	4,60	4,58	4,56	4,53	4,50	4,46	4,43	4,40	4,36	5
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,96	3,94	3,92	3,90	3,87	3,84	3,81	3,77	3,74	3,70	3,67	6
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,53	3,51	3,49	3,47	3,44	3,41	3,38	3,34	3,30	3,27	3,23	7
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,24	3,22	3,20	3,17	3,15	3,12	3,08	3,04	3,01	2,97	2,93	8
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,03	3,01	2,99	2,96	2,94	2,90	2,86	2,83	2,79	2,75	2,71	9
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,87	2,85	2,83	2,80	2,77	2,74	2,70	2,66	2,62	2,58	2,54	10
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,74	2,72	2,70	2,67	2,65	2,61	2,57	2,53	2,49	2,45	2,40	11
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,64	2,62	2,60	2,57	2,54	2,51	2,47	2,43	2,38	2,34	2,30	12
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,55	2,53	2,52	2,48	2,46	2,42	2,38	2,34	2,30	2,25	2,21	13
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,48	2,46	2,44	2,41	2,39	2,35	2,31	2,27	2,22	2,18	2,13	14
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,42	2,40	2,39	2,35	2,33	2,29	2,25	2,20	2,16	2,11	2,07	15
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,37	2,35	2,33	2,30	2,28	2,24	2,19	2,15	2,11	2,06	2,01	16
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,34	2,31	2,29	2,26	2,23	2,19	2,15	2,10	2,06	2,01	1,96	17
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,29	2,27	2,25	2,22	2,19	2,15	2,11	2,06	2,02	1,97	1,92	18
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,26	2,23	2,22	2,18	2,16	2,11	2,07	2,03	1,98	1,93	1,88	19
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,22	2,20	2,18	2,15	2,12	2,08	2,04	1,99	1,95	1,90	1,84	20
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,20	2,18	2,16	2,12	2,10	2,05	2,01	1,96	1,92	1,87	1,81	21
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,17	2,15	2,13	2,10	2,07	2,03	1,98	1,94	1,89	1,84	1,78	22
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,15	2,13	2,11	2,08	2,05	2,01	1,96	1,91	1,86	1,81	1,76	23
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,13	2,11	2,09	2,05	2,03	1,98	1,94	1,89	1,84	1,79	1,73	24
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,11	2,09	2,07	2,04	2,01	1,96	1,92	1,87	1,82	1,77	1,71	25
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,09	2,07	2,05	2,02	1,99	1,95	1,90	1,85	1,80	1,75	1,69	26
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,08	2,06	2,04	2,00	1,97	1,93	1,88	1,84	1,79	1,73	1,67	27
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,06	2,04	2,02	1,99	1,96	1,91	1,87	1,82	1,77	1,71	1,65	28
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,05	2,03	2,01	1,97	1,94	1,90	1,85	1,81	1,75	1,70	1,64	29
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,04	2,01	1,99	1,96	1,93	1,89	1,84	1,79	1,74	1,68	1,62	30
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,95	1,92	1,90	1,87	1,84	1,79	1,74	1,69	1,64	1,58	1,51	40
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,86	1,84	1,81	1,78	1,75	1,70	1,65	1,59	1,53	1,47	1,39	60
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,77	1,75	1,72	1,69	1,66	1,61	1,55	1,50	1,43	1,35	1,25	120
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,69	1,67	1,63	1,60	1,57	1,52	1,46	1,39	1,32	1,22	1,00	∞
	1	2	3	4	5	6	7	8	9	10	12	14	15	16	18	20	24	30	40	60	120	∞	

Exemplo 2. Os resultados da tabela abaixo são relativos às propriedades soporíferas da hiosciamina (droga A) e hioscina (droga B). Dois grupos de 10 pacientes são aleatoriamente selecionados e cada grupo toma uma das drogas. Os resultados em horas extras de sono são:

A	1,9	0,8	1,1	0,1	-0,1	4,4	5,5	1,6	4,6	3,4
B	0,7	-1,6	-0,2	-1,2	-0,1	3,4	3,7	0,8	0,0	2,0

Testar $H_0: \sigma_A^2 = \sigma_B^2$ vs. $H_1: \sigma_A^2 \neq \sigma_B^2$, ao nível de significância de 10%.

Solução:

$H_0: \sigma_A^2 = \sigma_B^2$

$H_1: \sigma_A^2 \neq \sigma_B^2$

$s_A^2 = 4{,}01$ $\quad$ $n_A = n_B = 10$ $\quad$ $\alpha = 10\%$

$s_B^2 = 3{,}20$

$$F_{obs} = \frac{s_A^2}{s_B^2} = \frac{4{,}01}{3{,}20} = 1{,}25, \quad gl = (9,9)$$

$F_{c\,(0,05;\,9,\,9)} = 3{,}18$ $\qquad$ $RC = \{F > 3{,}18\}$

Como $F_{obs} \notin RC$, não se rejeita H_0, ou seja, as variâncias são estatisticamente iguais ao nível de 10%.

A análise da hipótese da igualdade de variâncias é crucial para o uso do **teste t**, na comparação de duas médias, apresentado a seguir.

10.2 Comparação de duas médias de populações normais: amostras independentes

Com o objetivo de se comparar duas populações ou, sinonimamente, dois tratamentos, examinaremos a situação na qual os dados estão na forma de realizações de amostras aleatórias de tamanhos n_1 e n_2, selecionadas, respectivamente, das populações 1 e 2. Os dados são as medidas das respostas associadas com o seguinte delineamento experimental. Uma coleção de $n_1 + n_2$ elementos são aleatoriamente divididos em 2 grupos de tamanhos n_1 e n_2, onde cada membro do primeiro grupo recebe o tratamento 1 e do segundo, o tratamento 2. Especificamente, estaremos interessados em fazer inferência sobre o parâmetro:

(média da população 1) – (média da população 2) = $\mu_1 - \mu_2$

Formalmente, suponha uma amostra $X_1, X_2, \ldots, X_{n_1}$ selecionada aleatoriamente de uma população $N(\mu_1, \sigma_1^2)$ e uma amostra $Y_1, Y_2, \ldots, Y_{n_2}$ selecionada de uma população

N(μ_2, σ_2^2), n_1 e n_2 **independentes**. Para cada uma delas, teremos os respectivos estimadores da média e variância: $\bar{X}$ e S_1^2 e $\bar{Y}$ e S_2^2.

Hipótese: H_0 : $\mu_1 = \mu_2$ ou $\mu_1 - \mu_2 = 0$

Definindo a variável ($\bar{X} - \bar{Y}$), note-se que:

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2 \quad \text{e}$$

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) - 2Cov(\bar{X}, \bar{Y})$$

Como as variáveis $\bar{X}$ $e\bar{Y}$ são independentes, $Cov(\bar{X}, \bar{Y}) = 0$, então

$$Var(\bar{X} - \bar{Y}) = \sigma_1^2 / n_1 + \sigma_2^2 / n_2$$

Portanto, $(\bar{X} - \bar{Y})$ *tem distribuição* $N[(\mu_1 - \mu_2), (\sigma_1^2 / n_1 + \sigma_2^2 / n_2)]$

e, consequentemente,

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}} \tag{1}$$

tem distribuição N(0, 1).

10.2.1 1º caso: variâncias σ_1^2 e σ_2^2 conhecidas

Para testar a hipótese H_0 usa-se a estatística **(1)**. Como H_0 estabelece que $\mu_1 - \mu_2 = 0$,

$$Z = \frac{(\bar{X} - \bar{Y})}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

Hipóteses alternativas:	**Regiões críticas (nível α):**
$H_1 : \mu_1 \neq \mu_2$ ou $\mu_1 - \mu_2 \neq 0$	$Z > z_{c\,(\alpha/2)}$ ou $Z < -z_{c\,(\alpha/2)}$
$H_1 : \mu_1 > \mu_2$ ou $\mu_1 - \mu_2 > 0$	$Z > z_{c\,(\alpha)}$
$H_1 : \mu_1 < \mu_2$ ou $\mu_1 - \mu_2 < 0$	$Z < -z_{c\,(\alpha)}$

10.2.2 2º caso: variâncias desconhecidas e iguais

Preliminarmente, testa-se se as variâncias das duas populações são iguais. Caso a hipótese não seja rejeitada, isto é, que $\sigma_1^2 = \sigma_2^2 = \sigma^2$, a estatística **(1)** transforma-se em:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

Substituindo σ por um estimador, teremos uma expressão muito semelhante à t de Student. Uma estatística para σ^2 é a média ponderada:

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)},$$

que, como S_1^2 *e* S_1^2 são dois estimadores não viciados de σ^2, também é um estimador não viciado de σ^2.

O desvio padrão da diferença $(\overline{X} - \overline{Y})$ é estimado por:

$$S(\bar{X} - \bar{Y}) = S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

de modo que pode-se construir a estatística

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}}$$

que tem distribuição t de Student, com $n_1 + n_2 - 2$ graus de liberdade.

Sob H_0 verdadeira ($\mu_1 - \mu_2 = 0$), $t = \dfrac{(\bar{X} - \bar{Y})}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

Hipóteses alternativas:	**Regiões críticas (nível α):**
$H_1: \mu_1 \neq \mu_2$	$\lvert t \rvert > t_{c(\alpha/2,\ n_1 + n_2 - 2)}$
$H_1: \mu_1 > \mu_2$	$t > t_{c(\alpha,\ n_1 + n_2 - 2)}$
$H_1: \mu_1 < \mu_2$	$t < -t_{c(\alpha,\ n_1 + n_2 - 2)}$

Nota: quando ambas as amostras (n_1 e n_2) são pequenas ($n < 30$), o teste pode ser usado supondo, além da normalidade das distribuições das populações originais, que suas variâncias, σ_1^2 e σ_2^2, são iguais.

Exemplo 3. Usando os dados do exemplo 2, testar se há evidência de que as duas drogas são igualmente eficientes (H_0: $\mu_A = \mu_B$ vs. H_1: $\mu_A > \mu_B$), ao nível de 5%.

Solução:

H_0: $\mu_A = \mu_B$
H_1: $\mu_A > \mu_B$

$$n_A = n_B = 10 \qquad \bar{x}_A = 2{,}33 \qquad \bar{x}_B = 0{,}75$$
$$\alpha = 5\% \qquad s_A^2 = 4{,}01 \qquad s_B^2 = 3{,}20$$

$$s_P^2 = \frac{9.3{,}20 + 9.4{,}01}{18} = 3{,}61$$

$$t_{obs} = \frac{\bar{x}_A - \bar{x}_B}{s_P\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{2{,}33 - 0{,}75}{1{,}90\sqrt{\frac{1}{10} + \frac{1}{10}}} = 1{,}86$$

$t_{c(18;\,0,05)} = 1{,}734$ RC $= \{t > 1{,}734\}$

Como $t_{obs} \in$ RC, rejeita-se H_0, ou seja, há evidência de que a droga A é mais eficiente do que a B como soporífero.

10.2.3 3º caso: variâncias desconhecidas e desiguais (Teste de Smith – Satterthwaite)

Quando a hipótese de igualdade de variâncias for rejeitada, deve-se substituir σ_1^2 e σ_2^2 em **(1)** pelos seus respectivos estimadores, s_1^2 e s_2^2, obtendo a estatística:

$$t = \frac{(\bar{X} - \bar{Y})}{\sqrt{(s_1^2/n_1 + s_2^2/n_2}}$$

que, sob a veracidade de H_0 ($\mu_1 - \mu_2 = 0$), **aproxima-se** de uma distribuição t de Student, com número de graus de liberdade dado aproximadamente por:

$$gl = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Como o número de graus de liberdade assim calculado, geralmente, é **não inteiro**, recomenda-se aproximá-lo para o **inteiro** imediatamente anterior a este.

Se n_1 e n_2 são ambos grandes ($n \geq 30$), o teste pode ser baseado na estatística

$$Z = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{s_1^2/n_1+s_2^2/n_2}} \sim N(0,1) \quad \text{sob } H_0,$$

pois (1) permanece válido se σ_1^2 e σ_2^2 são substituídos por seus respectivos estimadores amostrais, s_1^2 e s_2^2.
A escolha da região de rejeição, mono ou bilateral, depende do tipo da hipótese alternativa.

Nota: no caso da inferência originada de amostras grandes, não é necessário assumir que as distribuições das populações originais são normais, porque o teorema limite central garante que as médias amostrais $\bar{X}$ e $\bar{Y}$ são aproximadamente distribuídas como $N(\mu_1, \sigma_1/\sqrt{n_1})$ e $N(\mu_2, \sigma_2/\sqrt{n_2})$, respectivamente. Além disso, a suposição de variâncias populacionais iguais $(\sigma_1^2 = \sigma_2^2)$, que é usada para amostras pequenas, é evitada nessa situação.

Exemplo 4. Querendo comparar o ganho em peso de duas raças de bovinos, A e B, num mesmo regime alimentar, tomaram-se n = 35 animais da raça A e m = 40 animais da raça B. Os resultados obtidos foram:

Raça	$\bar{x}$	s^2
A	70,5	81,6
B	84,3	200,5

Testar ao nível de 5% , se o ganho em peso médio das duas raças é o mesmo, ou seja H_0: $\mu_A = \mu_B$ vs. H_1: $\mu_A \neq \mu_B$.

Solução:

H_0: $\mu_A = \mu_B$
H_1: $\mu_A \neq \mu_B$

$n_A = 35$ $\quad$ $n_B = 40$ $\quad$ $\alpha = 5\%$

$$z_{obs} = \frac{(\bar{x}_B - \bar{x}_A)}{\sqrt{s_A^2/n_A + s_B^2/n_B}} = \frac{84{,}3 - 70{,}5}{\sqrt{81{,}6/35 + 200{,}5/40}} = \frac{13{,}8}{2{,}71} = 5{,}09$$

$z_c = 1{,}96$ $\quad$ RC = $\{z < -1{,}96$ ou $z > 1{,}96\}$

Como $z_{obs} \in$ RC, rejeita-se H_0, ou seja, há evidência que as duas raças têm ganhos em peso médios diferentes ($\mu_B > \mu_A$), ao nível de 5%.

10.3 Comparação emparelhada

Quando as médias de duas populações são comparadas, pode ocorrer uma diferença significativa entre elas por causa de fatores externos não controláveis, mesmo não havendo diferenças nos tratamentos avaliados. Reciprocamente, fatores externos podem mascarar ou ocultar uma diferença real. Uma maneira de contornar estes problemas é coletar as observações em pares, de modo que os dois elementos de cada par sejam homogêneos em todos os sentidos (por exemplo, quanto ao sexo, a idade, semelhança genética e de ambiente, etc.), exceto no que diz respeito aos tratamentos que se quer comparar. Assim, se houver uma diferença na resposta entre os dois grupos, esta pode ser atribuída a uma diferença nos tratamentos.
Tal planejamento é chamado **comparação emparelhada** e consiste em formarem pares e sortear os tratamentos dentro de cada par.

Como na formulação geral de comparação de duas médias, têm-se duas amostras $X_1, X_2,..., X_n$ e $Y_1, Y_2,..., Y_n$, só que agora as observações estão emparelhadas, isto é, a amostra é formada pelos pares $(X_1, Y_1),(X_2, Y_2),...,(X_n, Y_n)$.
Se definirmos a variável
$D_i = X_i - Y_i, \quad i = 1, 2 ,... , n$

teremos um conjunto de n observações, cada uma das quais é a diferença entre duas observações originais.

Os pares de observações $(X_i - Y_i)$ são independentes, mas X_i e Y_i dentro do i-ésimo par, são, geralmente, dependentes. Assim, se o emparelhamento das unidades experimentais for eficiente, espera-se X_i e Y_i ser, ao mesmo tempo, pequenos ou grandes, ou seja, ter uma correlação positiva alta. Um modo de se detectar isto é verificar se X e Y tem uma covariância positiva. Como

$$Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y),$$

a variância da diferença será menor neste caso do que seria no caso de variáveis aleatórias independentes, onde Cov(X, Y) = 0.
Esse procedimento também é usado quando as observações das duas amostras são feitas no mesmo indivíduo, por exemplo, medindo uma característica do indivíduo antes e depois dele ser submetido a um tratamento.
A estrutura das observações em uma comparação emparelhada é dada a seguir, onde X e Y denotam as respostas aos tratamentos 1 e 2, respectivamente.

	Tratamento		
Par	**1**	**2**	**Diferença (D_i)**
1	X_1	Y_1	$D_1 = X_1 - Y_1$
2	X_2	Y_2	$D_2 = X_2 - Y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	X_n	Y_n	$D_n = X_n - Y_n$

Definida as diferenças $D_i = X_i - Y_i$, i = 1, 2 ,... , n, é razoável assumir que elas constituem uma amostra aleatória de uma população com média = μ_D e variância σ_D^2, onde μ_D representa a diferença média real dos efeitos de tratamento dentro de pares. De outro modo,

$E(D_i) = E(X_i - Y_i) = \mu_D$ e

$Var(D_i) = Var(X_i - Y_i) = \sigma_D^2$, i = 1, 2 ,... , n

Se $\mu_D = \mu_1 - \mu_2 = 0$, então os dois tratamentos podem ser considerados equivalentes. Uma diferença positiva ($\mu_D > 0$) significa que o tratamento 1 tem uma resposta média maior do que a do tratamento 2.

A hipótese a ser testada é: H_0: $\mu_1 = \mu_2$ ou $\mu_D = 0$.

Hipóteses alternativas:

$$\begin{cases} H_1 : \mu_1 > \mu_2 \ ou \ \mu_D > 0 \\ H_1 : \mu_1 < \mu_2 \ ou \ \mu_D < 0 \qquad (Tratamento\,1\,tem\ resposta\,média\,menor\,do\,que\,a\,do\,2) \\ H_1 : \mu_1 \neq \mu_2 \ ou \ \mu_D \neq 0 \qquad (Tratamentos\,1\,e\,2\,tem\,respostas\ médias\,diferentes) \end{cases}$$

Supondo D_i : N (μ_D, σ_D^2),

$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \overline{X} - \overline{Y}$ tem distribuição N (μ_D, σ_D^2/n)

Definindo $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$, a estatística

$$t = \frac{\overline{D} - \mu_D}{s_D / \sqrt{n}}$$ tem distribuição t de Student, com n –1 graus de liberdade.

Como H_0 estabelece que $\mu_D = 0$, a fórmula de t é apresentada como

$$\frac{\overline{D}}{s_D/\sqrt{n}},$$ que é a estatística a ser usada no teste.

Quando n é grande (≥ 30), a inferência pode ser baseada na distribuição N(0, 1) ou equivalentemente na distribuição t com infinitos graus de liberdade (gl).
Note que há n pares de observações e apenas n – 1 gl. Se as observações não forem emparelhadas, mas tratadas como dois grupos independentes, teremos (n – 1) + (n – 1) = 2(n – 1) gl. A diminuição do número de gl resulta em um valor maior para $t_{\alpha/2}$, o que torna necessário um maior valor para t_{obs} atingir o limite de significância. Deste modo, se a formação de pares não for justificável, o teste será menos sensível, ou seja, preferindo pares, corre-se o risco de alguma perda de poder, a qual resulta em um aumento na probabilidade de aceitar a hipótese nula quando é falsa (β). O aumento é insignificante,

todavia, se o número de pares é grande, digamos, maior do que 10. O nível de significância (α) não é afetado.
Com um emparelhamento eficaz, a redução na variância da diferença (X – Y), geralmente, mais do que compensa a perda de graus de liberdade.

Exemplo 5. Cinco operadores de certo tipo de equipamento laboratorial são treinados em equipamentos de duas marcas diferentes, A e B. Mediu-se o tempo que cada um deles gastou na realização de uma mesma tarefa, e os resultados foram:

	Operador				
Marca	1	2	3	4	5
A	80	72	65	78	85
B	75	70	60	72	78

Ao nível de 1%, poderíamos afirmar que a tarefa realizada no equipamento A demora mais do que no B ($\mu_A > \mu_B$)?

Solução:

$H_0: \mu_A = \mu_B$

$\mathbf{H_1: \mu_A > \mu_B}$

$D_i = 5, 2, 5, 6, 7 \Rightarrow \begin{array}{l} \overline{D} = 5{,}0 \\ s_D = 1{,}87 \end{array}$ $\qquad$ n = 5 $\qquad$ $\alpha = 1\%$

$$t_{obs} = \frac{\overline{D}}{s_D / \sqrt{n}} = \frac{5{,}0}{1{,}87/\sqrt{5}} = 5{,}98$$

$t_{c(0,01;\,4)}$ = 3,747 $\qquad$ **RC = {t > 3,74}**

Como $t_{obs} \in$ RC, rejeita-se H_0, ou seja, a tarefa realizada no equipamento A demora mais do que no B ao nível de 1%.

10.4 Comparação de duas proporções binomiais

Vejamos agora como comparar as proporções de incidência de uma particular característica em duas populações. A estrutura da inferência é:

Parâmetro: $p_1 - p_2$ (proporção na população 1 - proporção na população 2)

Proporções amostrais: $\hat{p}_1 = \frac{X}{n_1}$ e $\hat{p}_2 = \frac{Y}{n_2}$, onde X e Y correspondem aos números de elementos que possuem a característica nas amostras n_1 e n_2, selecionadas aleatoriamente, respectivamente, das populações 1 e 2; n_1 e n_2 **independentes**.

Consideremos a estatística $\hat{p}_1 - \hat{p}_2$, como ponto de partida, para fazer a inferência sobre $p_1 - p_2$. Como a média e a variância das proporções amostrais são:

$$E(\hat{p}_1) = p_1 \qquad\qquad E(\hat{p}_2) = p_2$$

$$Var(\hat{p}_1) = \frac{p_1(1-p_1)}{n_1} \qquad\qquad Var(\hat{p}_2) = \frac{p_2(1-p_2)}{n_2}$$

e dado que $\hat{p}_1 \; e \; \hat{p}_2$ são independentes, a média e a variância da diferença $\hat{p}_1 - \hat{p}_2$ são:

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 \qquad \text{e} \qquad Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Logo, $DP(\hat{p}_1 - \hat{p}_2) = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$

O primeiro resultado mostra que $\hat{p}_1 - \hat{p}_2$ é uma estimador não viciado de $p_1 - p_2$ Uma estimativa do desvio padrão (DP) pode ser obtida substituindo p_1 e p_2 dentro da raiz por, respectivamente, $\hat{p}_1 \; e \; \hat{p}_2$. Além disso, para n_1 e n_2 grandes, a estatística $(\hat{p}_1 - \hat{p}_2)$ tem distribuição aproximadamente normal, de modo que:

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \quad \text{é aproximadamente N(0, 1).}$$

Para testar H_0: $p_1 = p_2$ ou $p_1 - p_2 = 0$, denota-se por **p** a proporção populacional conjunta não especificada.

Sob H_0 verdadeira, a estatística $(\hat{p}_1 - \hat{p}_2)$ é aproximadamente distribuída como normal, com

$$E(\hat{p}_1 - \hat{p}_2) = 0 \qquad \text{e} \qquad DP(\hat{p}_1 - \hat{p}_2) = \sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

O parâmetro p é estimado envolvendo as informações das duas amostras, ou seja,

$$\hat{p} = \frac{X+Y}{n_1 + n_2} \quad \text{(estimativa conjunta)}$$

Assim, considerando n_1 e n_2 grandes, a estatística

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \quad \text{é aproximadamente N(0, 1).}$$

Dependendo de H_1, a região crítica mono ou bi-caudal (regra de decisão) pode ser construída em termos da aproximação normal (Z).

Exemplo 6. Em um estudo sobre a incidência de abortos naturais entre médicas anestesistas (1) e de outras especialidades (2), obtiveram-se os seguintes resultados:

	1	2	Totais
Gestações normais	23	52	75
Abortos naturais	14	06	20
Totais	37	58	95

Denotando as proporções populacionais de abortos naturais em (1) e (2) por p_1 e p_2, respectivamente, testar $H_o: p_1 = p_2$ vs. $H_1: p_1 \neq p_2$, ao nível de 1%.

Solução:

H_0: $p_1 = p_2$

H_1: $p_1 \neq p_2$

$$\hat{p}_1 = \frac{14}{37} = 0{,}378 \qquad \hat{p} = \frac{14+6}{95} = 0{,}21$$

$$\hat{p}_2 = \frac{6}{58} = 0{,}103$$

$$z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0{,}378 - 0{,}103}{\sqrt{0{,}21.0{,}79}\sqrt{\frac{1}{37} + \frac{1}{58}}} = \frac{0{,}275}{0{,}086} = 3{,}19$$

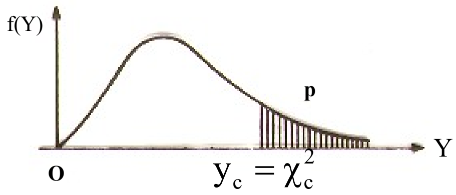
$\alpha = 1\%$ $\qquad z_c = 2{,}57 \qquad \Rightarrow \qquad RC = \{z > 2{,}57 \text{ ou } z < -2{,}57\}$

Como $z_{obs} \in RC$, rejeita-se H_0, ou seja, a proporção de abortos naturais em (1) é estatisticamente diferente (superior) da proporção em (2), ao nível de 1%.

Esse teste (Z) para comparações de duas proporções binomiais é equivalente ao teste qui-quadrado (χ^2) em uma tabela de contingência 2 x 2 (teste de homogeneidade de proporções), que será visto no próximo capítulo. Pode ser mostrado por cálculo algébrico que Z^2 é exatamente o mesmo que χ^2 para uma tabela assim especificada (2 x 2). Este é o caso do Exemplo 6, onde $Z^2 = \chi^2 \cong (3{,}19)^2 \cong 10{,}2$. Além disso, $(Z_{0,005} = 2{,}575)^2 = 6{,}63$ é o ponto crítico de $\chi^2(\chi^2_c)$, com $\alpha = 1\%$ e gl = 1. Entretanto, se o teste é monocaudal, tal como seria o caso com H_1: $p_1 > p_2$, o teste χ^2 não é apropriado.

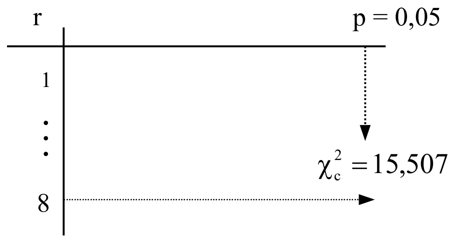
11 DISTRIBUIÇÃO QUI-QUADRADO

Seja Y uma variável aleatória contínua com distribuição qui-quadrado (χ^2) com **r** graus de liberdade. Graficamente, a distribuição χ^2 pode ser representada por:



$$P(Y > y_c) = p$$

Tal como no caso da distribuição t de Student, existe uma família de distribuições χ^2 indexada pelo número (inteiro) de graus de liberdade. A Tabela 6 fornece os valores de $y_c = \chi_c^2$ para alguns valores de **p** (α) e de **r** (graus de liberdade). Por exemplo,



Grau de liberdade (gl) é conceituado como o número de valores independentes de uma estatística, no caso, de χ^2, como será mostrado adiante.

11.1 Testes qui-quadrado

Serão apresentados aqui testes que utilizam a distribuição qui-quadrado como estrutura probabilística e por esta razão são denominados testes qui-quadrado. A figura acima apresenta a densidade do modelo χ^2 com a região crítica (RC) do teste, isto é, $RC = \{Y > \chi_c^2\}$.

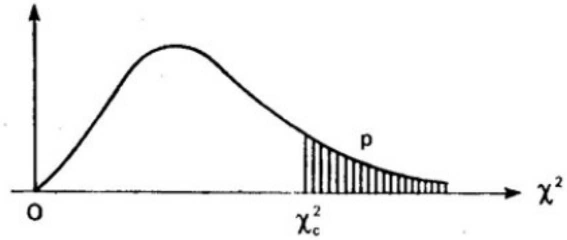
Esses testes são utilizados para dados discretos (categóricos) provenientes de uma população, tais como mortalidade ou achados patológicos, etc. O valor de qui-quadrado é um estimador da discrepância entre frequências esperadas e observadas, estabelecendo se as diferenças encontradas se devem ou não à casualidade.

TÁBUA IV

Distribuição de qui-quadrado: $\chi^2(n)$

Valores críticos de qui-quadrado tais que

$$P(\chi^2 > \chi_C^2) = p$$



Graus de liberdade n	p = 99%	98%	97,5%	95%	90%	80%	70%	50%	30%	20%	10%	5%	4%	2,5%	2%	1%	0,2%	0,1%	Graus de liberdade n
1	$0{,}0^{3}16$	$0{,}0^{3}63$	0,001	0,004	0,016	0,064	0,148	0,455	1,074	1,642	2,706	3,841	4,218	5,024	5,412	6,635	9,550	10,827	1
2	0,020	0,040	0,051	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	6,438	7,378	7,824	9,210	12,429	13,815	2
3	0,115	0,185	0,216	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	8,311	9,348	9,837	11,345	14,796	16,266	3
4	0,297	0,429	0,484	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	10,026	11,143	11,668	13,277	16,924	18,467	4
5	0,554	0,752	0,831	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	11,644	12,832	13,388	15,086	18,907	20,515	5
6	0,872	1,134	1,237	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	13,198	14,449	15,033	16,812	20,791	22,457	6
7	1,239	1,564	1,690	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	14,703	16,013	16,622	18,475	22,601	24,322	7
8	1,646	2,032	2,180	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	16,171	17,534	18,168	20,090	24,352	26,125	8
9	2,088	2,532	2,700	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	17,608	19,023	19,679	21,666	26,056	27,877	9
10	2,558	3,059	3,247	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	19,021	20,483	21,161	23,209	27,722	29,588	10
11	3,053	3,609	3,816	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	20,412	21,920	22,618	24,725	29,354	31,264	11
12	3,571	4,178	4,404	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	21,785	23,337	24,054	26,217	30,957	32,909	12
13	4,107	4,765	5,009	5,892	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	23,142	24,736	25,472	27,688	32,535	34,528	13
14	4,660	5,368	5,629	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	24,485	26,119	26,873	29,141	34,091	36,123	14
15	5,229	5,985	6,262	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	25,816	27,488	28,259	30,578	35,628	37,697	15
16	5,812	6,614	6,908	7,962	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	27,136	28,845	29,633	32,000	37,146	39,252	16
17	6,408	7,255	7,564	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	28,445	30,191	30,995	33,409	38,648	40,790	17
18	7,015	7,906	8,231	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	29,745	31,526	32,346	34,805	40,136	42,312	18
19	7,633	8,567	8,906	10,117	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	31,037	32,852	33,687	36,191	41,610	43,820	19
20	8,260	9,237	9,591	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	32,321	34,170	35,020	37,566	43,072	45,315	20
21	8,897	9,915	10,283	11,591	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	33,597	35,479	36,343	38,932	44,522	46,797	21
22	9,542	10,600	10,982	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	34,867	36,781	37,659	40,289	45,962	48,268	22
23	10,196	11,293	11,688	13,091	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	36,131	38,076	38,968	41,638	47,391	49,728	23
24	10,856	11,992	12,401	13,848	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	37,389	39,364	40,270	42,980	48,812	51,179	24
25	11,524	12,697	13,120	14,611	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	38,642	40,646	41,566	44,314	50,223	52,620	25
26	12,198	13,409	13,844	15,379	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	39,889	41,923	42,856	45,642	51,627	54,052	26
27	12,879	14,125	14,573	16,151	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	41,132	43,194	44,140	46,963	53,022	55,476	27
28	13,565	14,847	15,308	16,928	18,939	21,588	23,647	27,336	31,319	34,027	37,916	41,337	42,370	44,461	45,419	48,278	54,411	56,893	28
29	14,256	15,574	16,047	17,708	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	43,604	45,722	46,693	49,588	55,792	58,302	29
30	14,953	16,306	16,791	18,493	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	44,834	46,979	47,962	50,892	57,167	59,703	30
	p = 99%	98%	97,5%	95%	90%	80%	70%	50%	30%	20%	10%	5%	4%	2,5%	2%	1%	0,2%	0,1%	

11.2 Qui-quadrado como teste de aderência

O termo aderência refere-se à comparação de dados experimentais de frequência com a distribuição teórica.

Exemplo 2. Em ratos, o grupo sanguíneo Ag-B está associado a um lócus com vários alelos (alelos múltiplos), cuja segregação, em certos cruzamentos entre linhagens, parece apresentar desvios significativos de razões mendelianas. Os resultados (descendentes) do cruzamento entre as linhagens (heterozigotas) de ratos Ag-B^1Ag-B^4 x Ag-B^1Ag-B^4, foram:

Genótipos (k)	f_o	f_e sob H_0*
Ag-B^1Ag-B^1	58	50
Ag-B^1Ag-B^4	129	100
Ag-B^4Ag-B^4	13	50
Total (n)	200	200

* H_0 = a segregação segue a razão mendeliana 1 : 2 : 1

f_o = frequência observada

f_e = frequência esperada

que, à primeira vista, diferem da razão mendeliana 1 : 2 : 1. Formulando–se a hipótese H_0 de que a segregação é 1 : 2 : 1, as f_e's dos três genótipos são, respectivamente, 200.(1/4) = 50, 200.(2/4) = 100 e 200.(1/4) = 50.

Para testar se os números observados (f_o) dos três genótipos são consistentes com os esperados (f_e) com base na segregação 1 : 2 : 1, usa-se, então, a estatística:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_0 - f_e)^2}{f_e}$$

que sob H_0 tem distribuição χ^2 (qui-quadrado) com **r = k - 1** graus de liberdade.
Note que em **r**, se subtrai 1 de k por causa da condição de restrição que estabelece que, sendo conhecidas (k-1) frequências esperadas (independentes), a remanescente pode ser determinada por diferença.
Quando as f_e's somente puderem ser calculadas mediante estimativas de **m** parâmetros populacionais, a partir de estatísticas amostrais, o número de graus de liberdade (**r**) é dado por **r = k – 1 – m**.
Formalmente, fixado α, rejeita-se H_0 se $\chi^2 > \chi^2_{\alpha,r}$, onde $\chi^2_{\alpha,\, r}$ denota o ponto para o qual uma variável Y, distribuída como χ^2 com r graus de liberdade, satisfaz $P(Y > y_c) = \alpha$.
É importante notar que só se rejeita H_0 à medida que a frequência observada se afasta da esperada, ou seja, quando os valores obtidos para o χ^2 forem grandes.

11.2.1 Procedimento do teste:

1. Enunciar H_0 e H_1

$$\begin{cases} H_0 : \text{a segregação está de acordo com a razão mendeliana } 1:2:1 \\ H_1 : \text{a segregação é diferente de } 1:2:1 \end{cases}$$

2. Fixar α (nível de significância)

3. Calcular χ^2_{obs}

$$\chi^2_{obs} = \frac{(58-50)^2}{50} + \frac{(129-100)^2}{100} + \frac{(13-50)^2}{50} = 1{,}28 + 8{,}41 + 27{,}38 = 37{,}07$$

4. Determinar a região crítica

$$RC = \{\chi^2 > \chi^2_{c(\alpha,\, k-1)}\}$$

como $k - 1 = 2$ e se $\alpha = 1\% \Rightarrow \chi^2_c = 9{,}21$

5. Estabelecer a regra de decisão

Rejeitar H_0 se $\chi^2_{obs} \geq \chi^2_c$

6. Concluir

Como $\chi^2_{obs} > \chi^2_c$, rejeita-se H_0 (a hipótese que os resultados estão de acordo com a razão mendeliana 1 : 2 : 1).

Exemplo 3. Seja **t** o número eventual de hemáceas presentes em um volume representado pelo pequeno quadrado observado em um hemocitômetro. Sendo $\mathbf{f_o}$ a frequência observada, suponha o seguinte resultado:

t	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
$\mathbf{f_o}$	0	0	1	3	5	10	15	20	17	6	3	0	0	**80**
$\mathbf{t.f_o}$	0	0	2	9	20	50	90	140	136	54	30	0	0	**531**

Testar se o modelo de Poisson descreve adequadamente os dados da tabela.

Solução:

$$\hat{\lambda} = \sum t \,.\, f_o \,/ \sum f_o = 531/80 = 6{,}6 \qquad\qquad P(X = t) = \frac{\lambda^t e^{-\lambda}}{t!} = \frac{6{,}6^t\, e^{-6{,}6}}{t!}$$

Fazendo t = 4,

$$P(X = 4) = \frac{(6{,}6)^4\, e^{-6{,}6}}{4!} \approx 0{,}11$$

e a frequência esperada por Poisson é : $P(X = 4) \,.\sum f_o = 0{,}11.80 \approx 9{,}0$

Assim procedendo,

t	≤3	4	5	6	7	8	9	10	≥11	Total
f_o	4	5	10	15	20	17	6	3	0	80
f_e por Poisson	8	9	11	13	12	10	7	5	5	80

As frequências esperadas das três primeiras classes de **t** e das duas últimas são menores do que 5. Como a validade do teste de aderência, exclui essa situação, as três primeiras classes foram combinadas com a posterior (quarta) e as duas últimas combinadas entre si. A estatística χ^2 e o número de graus de liberdade são, então, calculados a partir dessas classes convenientemente modificadas.

H_0: dados são distribuídos segundo Poisson

$$\chi^2_{obs} = \sum_{\substack{todas \\ as\ céls.}} \frac{(f_0 - f_e)^2}{f_e} = 14{,}7$$

gl = nº de classes (9) - 1 - nº de parâmetros estimados [1 (λ)] = 7

$$\chi^2_{c\,(1\%,\,7)} = 18{,}48$$

Portanto, como $\chi^2_{obs} < \chi^2_c$, não há evidência suficiente para se rejeitar a hipótese de que os dados são distribuídos segundo Poisson.

11.3 Teste qui - quadrado em tabelas de contingência

A classificação de observações (em geral, de variáveis qualitativas) de acordo com dois critérios é referida como **tabela de contingência**.

Exemplo 4. Natureza de vacas, segundo a raça e o tipo de acasalamento

	Tipo de acasalamento		
Raça	Fecundos	Não-fecundos	**Total**
Charolesa	110 (120)	50 (40)	160
Gir	70 (60)	10 (20)	80
Nelore	30 (30)	10 (10)	40
Total	210	70	280

Se um critério envolve **m** categorias (linhas) e o outro **n** categorias (colunas), a tabela é referida como tabela **m × n**. No exemplo, a tabela é **3 × 2**.

Tabelas de contingência são construídas com o propósito de se testar:

(1) a relação de dependência (associação) entre duas variáveis (**Teste de independência**). O teste de independência é baseado no esquema amostral, no qual uma única amostra aleatória de tamanho **n** é classificada com relação a duas características simultaneamente;

(2) que as várias colunas (ou linhas) tem a mesma proporção de indivíduos nas várias categorias de uma característica, se os totais das linhas (ou colunas) são especificados antecipadamente (**Teste de homogeneidade**).

11.3.1 Teste de homogeneidade

Utilizando o Exemplo 4, iremos testar a igualdade das proporções de acasalamentos fecundos (e não fecundos) nas três raças. Vejamos os passos a seguir:

1. Estabelecer H_0 e H_1

A hipótese nula de homogeneidade que a proporção de cada tipo de acasalamento é a mesma para todas as raças, pode ser formalmente estabelecida como:

$$H_0: p_{Ch(j)} = p_{Gir(j)} = p_{Ne(j)} \quad \text{para cada } j = 1 \text{ (fecundo) e } 2 \text{ (não fecundo)}$$

Ou simplesmente,

$$\begin{cases} H_0: \text{ a proporção de acasalamentos fecundos é a mesma nas três} \\ \quad \text{raças ou seja, } p_{Ch} = p_{Gir} = p_{Ne}. \textit{ Assim,} \\ H_1: \text{ as proporções não são todas iguais.} \end{cases}$$

2. Calcular as f_e's sob a hipótese H_0 ser verdadeira

Dos 280 animais → 210 fecundos

Dos 160 Charolês → X fecundos $\qquad X = \dfrac{160.210}{280} = 120$

Analogamente,

Dos 280 animais → 210 fecundos

Dos 80 Gir → X fecundos $\qquad X = \dfrac{80.210}{280} = 60$

Todas as demais f_e's podem ser calculadas por diferença (os valores calculados estão entre parênteses na tabela). Diz-se então que há 2 graus de liberdade. Isso corresponde a (m – 1) . (n – 1) graus de liberdade, ou seja:

$r = (m – 1) . (n – 1) = (3 – 1) . (2 – 1) = 2$

Este procedimento pode ser interpretado como: dados os totais marginais, calcula-se que números seriam esperados na tabela a fim de tornarem as proporções de fecundidade para as três raças exatamente iguais. Assim, na célula da 1ª linha e 1ª coluna esse número esperado é (210/280) .160 = 120,
já que a proporção de fecundidade geral é 210/280 e há 160 indivíduos na
raça Charolesa. Prosseguindo-se dessa forma obtêm-se os demais números esperados.

3. Calcular o valor da estatística

$$\chi^2_{obs} = \sum_m \sum_n \frac{(f_0 - f_e)^2}{f_e} = \frac{(110-120)^2}{120} + \ldots + \frac{(10-10)^2}{10} = 9{,}99$$

4. Determinar a região crítica

com gl = (m – 1) . (n – 1) = (2) . (1) = 2 e $\alpha = 5\% \Rightarrow \chi^2_c = 5{,}99$

RC = $\{\chi^2 > 5{,}99\}$

5. Estabelecer a regra de decisão

Rejeitar H_0 se $\chi^2_{obs} \geq \chi^2_c = 5{,}99$

6. Concluir

Como $\chi^2_{obs} > \chi^2_c$, rejeita-se H_0 ou seja, as fecundidades das raças não são todas estatisticamente iguais, ao nível de 5%.

Como H_0 foi rejeitada, deve-se continuar a investigação, comparando-se as raças duas a duas, para se verificar quem difere de quem em termos do critério analisado.

11.3.2 Tabela de contingência 2 × 2 (comparação de duas proporções)

Exemplo 5. Considerando a seguinte tabela:

Tratamento	Morte	Sobrevivência	**Total**
A	41 (53,86)	216 (203,14)	**257**
B	64 (51,14)	180 (192,86)	**244**
Total	**105**	**396**	**501**

verificar se os dados proporcionam evidência que as proporções de mortalidade são diferentes para os dois tratamentos ($\alpha = 1\%$).

Solução:

$H_0 : p_A = p_B$
$H_1 : p_A \neq p_B$

em que: p_A e p_B denotam as proporções de morte (ou de sobrevivência) para os tratamentos A e B, respectivamente.

$f_{e(}$= (105.257)/501 = 53,86

e as demais por diferença (valores entre parênteses na tabela) gl = (2 – 1) . (2 –1) = 1

$$\chi^2_{obs} = \frac{(41-53{,}86)^2}{53{,}86} + ... + \frac{(180-192{,}86)^2}{192{,}86} = 7{,}97$$

$$\chi^2_{c(1\%;1)} = 6{,}63$$

Como $\chi^2_{obs} > \chi^2_c$, rejeita-se H_0, ou seja, há uma diferença real entre as proporções de mortalidade (ou de sobrevivência) provocada pelos tratamentos A e B.

Para tabelas de contingência 2 x 2, o valor de χ^2 pode ser obtido também pela fórmula (1):

			Total
	a	b	n_1
	c	d	n_2
Total	n_3	n_4	N

$$\chi^2_{obs} = \frac{(c.b - a.d)^2.N}{n_1.\,n_2.\,n_3.\,n_4} \qquad (1)$$

Então, $\chi^2_{obs} = \dfrac{(216.64 - 41.180)^2.501}{(257).(244).(105).(396)} = \dfrac{(13.824 - 7.380)^2.501}{2.607.398.640} = 7{,}97$

Nas tabelas de contingência 2 x 2, alguns autores recomendam usar o teste de χ^2 com a **correção de Yates** para continuidade. Esta correção consiste em subtrair ½ de cada diferença (f_o – f_e) antes de elevá-la ao quadrado. Com este procedimento a fórmula (1) transforma-se em:

$$\chi^2_{obs} = \frac{(c.b - a.d - \frac{N}{2})^2.N}{n_1.\,n_2.\,n_3.\,n_4}$$

Com a correção de Yates, o valor de χ^2 no Exemplo 5 torna - se 7,37, mostrando que em amostras grandes, produz, praticamente, o mesmo resultado que o χ^2 não corrigido. A correção tem importância principalmente quando os valores das f_e's são pequenos, mas se a menor f_e for < 5, deve-se, então, usar o **teste exato de Fisher**, que é baseado exclusivamente no cálculo de probabilidades. Não trataremos, entretanto, deste teste.

Obs. Pode ser mostrado por cálculo algébrico que Z^2 $\left(Z = \dfrac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{(1/n_A)+(1/n_B)}} \right)$ é exatamente o mesmo que χ^2 para uma tabela de contingência 2 x 2. Este é o caso do Exemplo 5, onde: $Z^2 = \chi^2 \cong 7{,}97$. Além disso, $(Z_{0{,}005} = 2{,}575)^2 = 6{,}63$ é o ponto crítico de χ^2 (χ^2_c), com α = 1% e gl = 1. Assim, esses dois testes são equivalentes para comparação de duas proporções. Entretanto, se o teste é monocaudal, tal como é o caso com H_1: $p_1 > p_2$, o teste χ^2 não é apropriado.

Teste de independência

O procedimento para o **teste de independência** é equivalente ao apresentado para o **teste de homogeneidade,** ou seja, as fórmulas para χ^2 e graus de liberdade são os mesmos tanto para o teste de homogeneidade como para o de independência. Somente o método amostral e a formalização de H_0 são diferentes para as duas situações.
Para um tratamento geral do **teste de independência** em uma tabela de contingência r x c, suponha n indivíduos classificados de acordo com dois critérios: A e B, e que há r categorias para A ($A_1, A_2, ..., A_r$) e c categorias para B ($B_1, B_2, ..., B_c$). Colocando a categoria A nas linhas e B nas colunas, pode-se construir uma tabela de dupla entrada, na qual cada célula é a intersecção de A com B.
A hipótese nula que se interessa testar é que as classificações A e B são independentes. Relembrando que a probabilidade da intersecção de eventos independentes é o produto de suas probabilidades, logo a hipótese nula de independência, estabelecendo que os eventos $A_1, A_2, ..., A_r$ são independentes dos eventos $B_1, B_2, ..., B_c$, pode ser representada por : $P(A_iB_j) = P(A_i).P(B_j)$. Ou seja, numa tabela de contingência de **r** linhas e **c** colunas, a hipótese nula de independência é:

$$\mathbf{H_0 : p_{ij} = p_{i.} \cdot p_{.j}} \qquad \text{para todo} \quad \begin{bmatrix} i = 1,2,\ldots,r \\ j = 1,2,\ldots,c \end{bmatrix}$$

Em outras palavras, fazendo $\mathbf{p_{ij}}$, a probabilidade de um indivíduo, selecionado ao acaso, pertencer à célula da linha **i** e da coluna **j**, $\mathbf{p_{i.}}$, a probabilidade dele pertencer à linha **i** (total marginal) e $\mathbf{p_{.j}}$, a probabilidade de pertencer à coluna **j** (total marginal), têm-se que as probabilidades no corpo da tabela ($\mathbf{p_{ij}}$) serão os produtos dos totais marginais ($\mathbf{p_{ij} = p_{i.} \cdot p_{.j}}$), se os critérios **i** e **j** forem **independentes**.

No caso do exemplo 5, se os eventos A e M, correspondentes ao tratamento A e a ocorrência de morte, respectivamente, forem independentes, $P(A \cap M) = P(A).P(M) = \frac{257}{501} \cdot \frac{105}{501} = 0{,}1075$. Assim, na célula da 1ª linha e 1ª coluna, o número esperado é $0{,}1075.501 = 53{,}86 = \frac{257.105}{501}$, tal como no teste de homogeneidade.
Prosseguindo dessa forma ou por diferença, obtêm-se os demais números esperados
Exemplo 6. Teste de independência entre os atributos sexo e grupo sanguíneo, considerando uma amostra de 367 indivíduos, classificados de acordo com as duas características simultaneamente.

	Grupo sanguíneo				
Sexo	O	A	B	AB	Total
Masculino	96(99*)	94(98)	30(24)	14(13)	234
Feminino	59(56)	60(56)	7(13)	7(8)	133
Total	155	154	37	21	367

Os valores entre parênteses na tabela correspondem às frequências esperadas calculadas sob a hipótese H_0 ser verdadeira [* = (155.234)/367]

H_0: os dois atributos são independentes
H_1: os dois atributos não são independentes

$$\chi^2_{obs} = \frac{(96-99)^2}{99} + \frac{(59-56)^2}{56} + \cdots + \frac{(7-8)^2}{8} = 5{,}2$$

$$\chi^2_{c(5\%,\ 3)} = 7{,}82$$

Conclusão: como $\chi^2_{obs} < \chi^2_{c(5\%,\ 3)}$, a hipótese de independência entre os dois atributos (sexo e grupo sanguíneo) não é rejeitada ao nível de significância de 5%.

Restrições do uso do teste qui-quadrado (χ^2)

Por razões teóricas:
- os testes vistos são aplicados sem restrição se todas as frequências esperadas forem maiores do que 5;
- quando o grau de liberdade for igual a 1, cada frequência esperada não deve ser inferior a 5;
- quando o grau de liberdade for maior do que 1, o teste qui-quadrado não deve ser usado se mais de 20% das frequências esperadas forem inferiores a 5 ou se qualquer frequência esperada for inferior a 1.
- os testes somente devem ser aplicados aos dados observados e nunca com as proporções ou porcentagens oriundas dos mesmos.

Obs.: caso haja restrições no uso do teste, eventualmente, pode-se juntar categorias adjacentes de modo a aumentar as frequências esperadas.

12 REGRESSÃO E CORRELAÇÃO LINEAR

12.1 . Introdução: regressão versus correlação

Em experimentos que procuram determinar a relação existente entre duas variáveis, por exemplo, a dose de uma droga e a reação, concentração e densidade ótica, peso e altura, idade da vaca e a produção de leite, etc., dois tipos de situações podem ocorrer:

(a) uma variável (X) pode ser medida acuradamente e seu valor escolhido pelo experimentador. Por exemplo, a dose de uma droga a ser ministrada no animal. Esta variável é a **variável independente**. A outra variável (Y), dita **variável dependente ou resposta**, está sujeita a erro experimental, e seu valor depende do valor escolhido para a variável independente. Assim, a resposta (reação, Y) é uma variável dependente da variável independente dose (X). Este é o caso da **Regressão**.

(b) as duas variáveis quando medidas estão sujeitas a erros experimentais, isto é, erros de natureza aleatória inerentes ao experimento. Por exemplo, produção de leite e produção de gordura medidas em vacas em lactação, peso do pai e peso do filho, comprimento e a largura do crânio de animais, etc. Este tipo de associação entre duas variáveis constitui o problema da **Correlação**.

Atualmente, se dá à técnica de correlação uma importância menor do que a da regressão. Se duas variáveis estão correlacionadas, é muito mais útil estudar as posições de uma ou de ambas por meio de curvas de regressão, as quais permitem, por exemplo, a predição de uma variável em função de outra, do que estudá-las por meio de um simples coeficiente de correlação.

12.2 Regressão linear simples

O termo regressão é usado para designar a expressão de uma variável dependente (Y) em função de outra (X), considerada independente. Diz-se regressão de Y em (sobre) X. Se a relação funcional entre elas é expressa por uma equação do 1º grau, cuja representação geométrica é uma linha reta, a regressão é dita linear.

Para introduzir a ideia de regressão linear simples, consideremos o seguinte exemplo:

Tabela 1. Tempo, em minutos, e quantidade de procaina[1] hidrolisada, em 10^{-5} moles/litro, no plasma canino.

	Tempo(X)	Quantidade hidrolisada (Y)	X.Y	X^2	Y^2
	2	3,5	7,0	4,0	12,3
	3	5,7	17,1	9,0	32,5
	5	9,9	49,5	25,0	98,0
	8	16,3	130,4	64,0	265,7
	10	19,3	193,0	100,0	372,5
	12	25,7	308,4	144,0	660,5
	14	28,2	394,8	196,0	795,2
	15	32,6	489,0	225,0	1062,8
Total	69	141,2	1589,2	767,0	3299,5

[1] anestésico local

A simples observação dos dados apresentados na Tabela 1 mostra que no intervalo estudado a quantidade de procaina hidrolisada varia em função do tempo.
Na resolução de problemas de regressão, o primeiro passo é traçar o **diagrama de dispersão** correspondente, marcando, em um sistema cartesiano bidimensional, os diversos pares de valores observados (x_i , y_i). Assim, o diagrama de dispersão correspondente aos dados da Tabela 1 é mostrado na Figura 1.

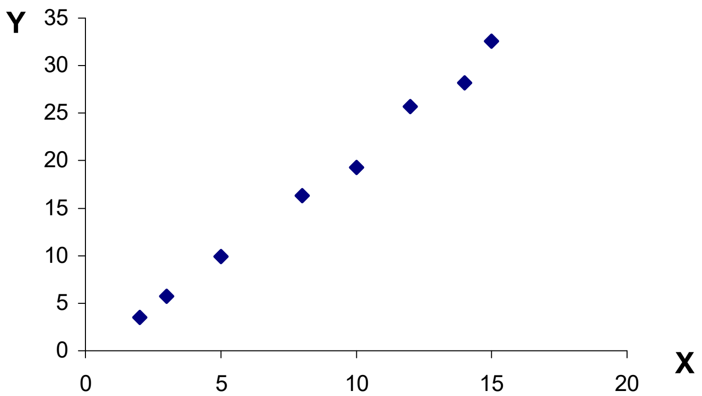


Figura 1. Diagrama de dispersão dos dados da Tabela 1.

É fácil ver observando essa figura, que os pontos relativos aos dados de tempo e quantidade de procaina hidrolisada estão praticamente sobre uma reta. Parece então razoável estabelecer que a variação da quantidade de procaina hidrolisada (Y) pode ser considerada como uma função linear do tempo (X).

Postulada a existência de uma relação linear entre duas variáveis, pode-se representar o conjunto de pontos (x_i, y_i) pela equação da reta:

$$y = \alpha + \beta x + \varepsilon$$

que expressa o valor de Y como função do valor de X, onde ε, conhecido como *erro* ou *resíduo*, é a distância que um resultado y em particular se encontra da linha de regressão da população, representada pela equação:

$$E(y/x) = \alpha + \beta x,$$

em que α indica o intercepto da linha com o eixo do Y e β o coeficiente angular ou inclinação da reta.
Se ε [y – E(y/x)] é positivo, y é maior do que E(y/x); se é negativo, y é menor do que E(y/x); e a soma dos $\varepsilon_i^{'}s$ é igual a zero ($\sum \varepsilon_i = 0$). Logo, a média dos erros é nula, isto é, $E(\varepsilon_i) = 0$.
Como veremos a seguir, os parâmetros α e β da linha de regressão da população são estimados a partir da amostra aleatória de observações (x_i, y_i).

Regressão linear: estimação de parâmetros

Considerando, então, que observações $x_1, x_2, ..., x_k$ sejam obtidas sobre a variável independente x, tal que $y_1, y_2, ..., y_k$ sejam as observações feitas sobre a variável dependente y, todas sujeitas a erros experimentais, pode-se querer saber como é que y varia, em média, para um dado x. Ou seja, como os $y_s^{'}$ variam aleatoriamente, deseja-se conhecer a distribuição do y quando x é conhecido. Isto é feito por meio da **esperança condicionada de y dado x**, simbolizada por **E(y/x)**, que depende em geral de x. E(y/x) é também chamada de **função de regressão de y em x**.
A Figura 2, apresentada a seguir, mostra as distribuições de y dados certos valores de x, supondo a função de regressão de y em x linear.

Modelo. A reta da Figura 2 é simbolizada por $E(y/x) = \alpha + \beta x$, onde α e β são os parâmetros a serem estimados.
A partir de agora, se o modelo acima for desenvolvido num contexto paramétrico, uma hipótese simplificadora e muito simples deve ser feita, a saber: a distribuição da variável aleatória y, para um dado x, é normal. Mais especificamente, fixado um x_i (X não é uma variável aleatória), os $y_s^{'}$ constituem variáveis independentes normais $N(\alpha + \beta x_i, \sigma^2)$; o que equivale dizer que as médias das distribuições de y/x estão sobre a verdadeira reta $\alpha + \beta x$ ou seja, $E(y_i) = E(\alpha) + E(\beta x_i) + E(\varepsilon_i) = \alpha + \beta x_i$, onde $E(\varepsilon_i) = 0$, e que para um dado valor de x, a variância do erro é sempre σ^2, denominada variância residual, isto é, $E[y_i - E(y_i/x_i)]^2 = E(\varepsilon_i)^2 = \sigma^2$ (propriedade homocedástica). Estes conceitos estão ilustrados na Figura 2.

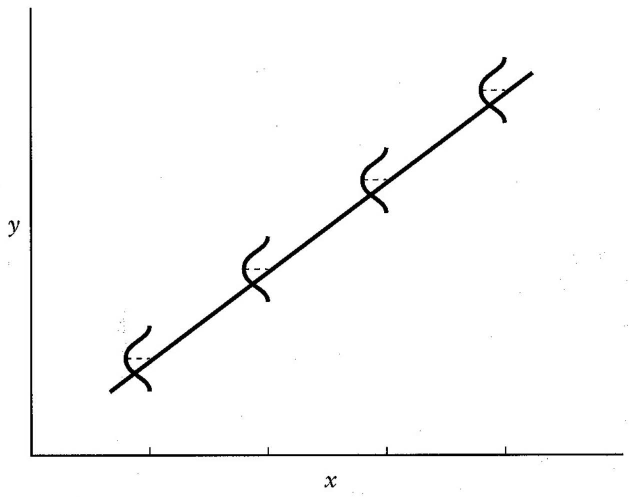


Figura 2. Normalidade dos resultados y para determinado valor de x

À parte do fato que σ^2 é desconhecido, a reta na qual as médias estão localizadas é também desconhecida. Assim, um objetivo importante da análise estatística é estimar os parâmetros α e β para que se conheça totalmente a função de regressão **E(y/x)**. A teoria mostra que a melhor maneira de estimá-los é por meio do **método dos quadrados mínimos**, que consiste em minimizar a soma dos quadrados das distâncias $\mathbf{y}_i - \hat{\mathbf{y}}_i$, onde $\hat{\mathbf{y}}_i = \mathbf{a} + \mathbf{b}\mathbf{x}_i$ representa a equação de regressão estimada, tal que $\mathbf{a} = \hat{\alpha}$ e $\mathbf{b} = \hat{\beta}$ são os estimadores de α e β, respectivamente.

Sendo, então, $y_i - \hat{y}_i$ a diferença entre o valor observado e o estimado pela equação de regressão para cada observação, a qual é rotulada por **e_i**, procura-se estimar α e β, de modo que $\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$ seja o menor possível. As diferenças e_i = $y_i - \hat{y}_i$ são chamadas "desvios da regressão" ou "erros de estimativas". Se todos os desvios (e_i) são iguais a zero, implica que cada ponto (x_i, y_i) se encontra diretamente sobre a linha ajustada; os pontos estão tão próximos quanto possíveis da linha.

Estimadores. Dado um conjunto de n pares de observações (x_1, y_1), (x_2, y_2), ... , (x_n, y_n), pode-se mostrar, usando métodos de cálculo infinitesimal não utilizado aqui, que os estimadores de quadrados mínimos são:

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

Dividindo-se o numerador e o denominador de b por (n − 1), vê-se que

$$b = \frac{Cov(X,Y)}{s_X^2} = \frac{[\sum (x_i - \bar{x})(y_i - \bar{y})]/n-1}{[\sum (x_i - \bar{x})^2]/n-1}$$

b é denominado coeficiente de regressão de Y em X; simboliza-se por $b_{Y.X}$

Fórmulas de cálculo:

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$\sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Note-se que, além da suposição da normalidade do y, outras hipóteses usadas pelo método de mínimos quadrados são:

para qualquer valor específico de x, $\sigma_{y/x}$, o desvio padrão dos resultados y, não se modifica. Esta hipótese de variabilidade constante em todos os valores de x é conhecida como *homoscedasticidade*, e

(b) a relação (verdadeira) entre y e x é suposta linear; mais claramente,
E(y/x) = α + βx.

Vejamos agora o cálculo da equação de regressão usando como exemplo os dados apresentados na Tabela 1:

$$b_{Y.X} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{1589{,}2 - \frac{69.141{,}2}{8}}{767 - \frac{(69)^2}{8}} = \frac{371{,}35}{171{,}88} = 2{,}16$$

$$a = \bar{y} - b\bar{x} = \frac{141{,}2}{8} - (2{,}16\,\frac{69}{8}) = 17{,}65 - (2{,}16\ .\ 8{,}63) = -\ 0{,}98$$

Portanto, a **equação de regressão linear** é:

$$\hat{y}_i = -0{,}98 + 2{,}16.x_i \qquad (1)$$

ou, como $a = \bar{y} - b\bar{x}$ e $\hat{y} = \bar{y} - b\bar{x} + bx$,

$$\hat{y}_i = \bar{y} + b(x_i - \bar{x}) \ = \mathbf{17{,}65 + 2{,}16\ (x_i - 8{,}63)} \qquad (2)$$

Note que as equações (1) e (2) são equivalentes; entretanto, em (2) fica mais evidente que a reta de regressão passa pelo ponto $(\bar{x}, \bar{y})$. O coeficiente angular da reta (b) é positivo, tal como sugerido pelo próprio diagrama de dispersão.
Para traçar a reta de regressão, basta dar valores quaisquer para X dentro do intervalo estudado e calcular os respectivos valores de $\hat{Y}$ (Figura 3). Os valores calculados de $\hat{Y}$ não coincidem necessariamente com os valores observados de Y. A curva resultante é denominada de regressão de Y para X, visto que Y é avaliado a partir de X.

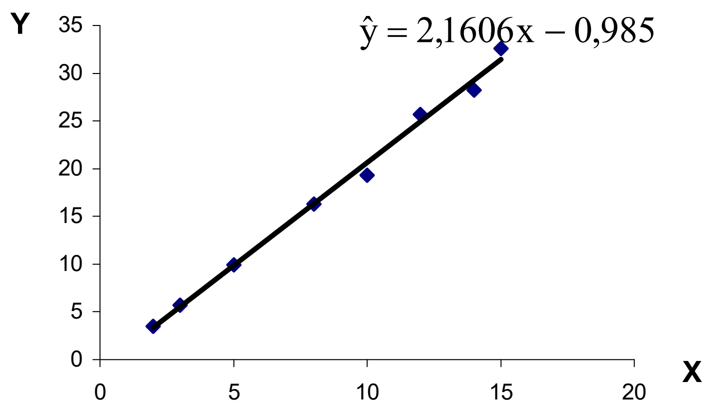


Figura 3. Quantidade de procaina hidrolisada ($\hat{Y}$) em função do tempo (X).

O mais importante objetivo de um estudo de regressão é usar o modelo linear desenvolvido para estimar a resposta esperada correspondente a um nível específico da variável controlada. De acordo com o modelo linear, a resposta esperada para um valor x da variável controlada é dada por $E(y/x) = \alpha + \beta\, x$ e a estimada, por $\hat{y} = a + bx$, que é um estimador não viciado para a média $E(y/x)$. Isto é, como pode ser mostrado, $E(\hat{y}/x) = E(a) + x\,E(b) = \alpha + \beta\, x$. Assim, por exemplo, na equação de regressão linear (1), para $x_i = 11'$, $\hat{y}_i = 22{,}8.10^{-5}$ moles/litro.

É importante aqui distinguir entre **interpolação** (predição dentro da amplitude dos dados amostrados; no exemplo, predição da quantidade de procaína hidrolisada no tempo igual há 11 minutos) e **extrapolação** (predição fora da amplitude dos dados; no exemplo, predição da quantidade de procaína hidrolisada no tempo de 17′ como sendo aproximadamente $35{,}7.10^{-5}$ moles/litro). A extrapolação deve-se implementada com cuidado, pois, (1) embora existindo uma relação linear entre X e Y (esta pode ser adequada na região definida pelo conjunto de valores usados), o modelo pode deixar de ser válido fora da região definida por esse conjunto, e (2) quanto mais afastado o valor predito (x_i) estiver de $\bar{x}$, maior será o erro da extrapolação.

12.3 Interpretação do coeficiente de regressão (b)

Obtida uma reta de regressão, o primeiro passo na sua interpretação é verificar o sinal de b. Se for positivo, indica que, quanto maior o valor de X, maior o valor de Y; se negativo, indica que quanto maior o valor de X, menor o valor de Y.

Uma interpretação mais informativa para o coeficiente de regressão (b) é que ele representa em quanto varia a média de Y para o aumento de uma unidade da variável X. Esta variação pode ser negativa, situação em que para um acréscimo de X corresponde um decréscimo de Y. Esse coeficiente, juntamente com o intercepto (a), o qual determina o ponto em que a reta corta o eixo de Y, estão representados na Figura 4.

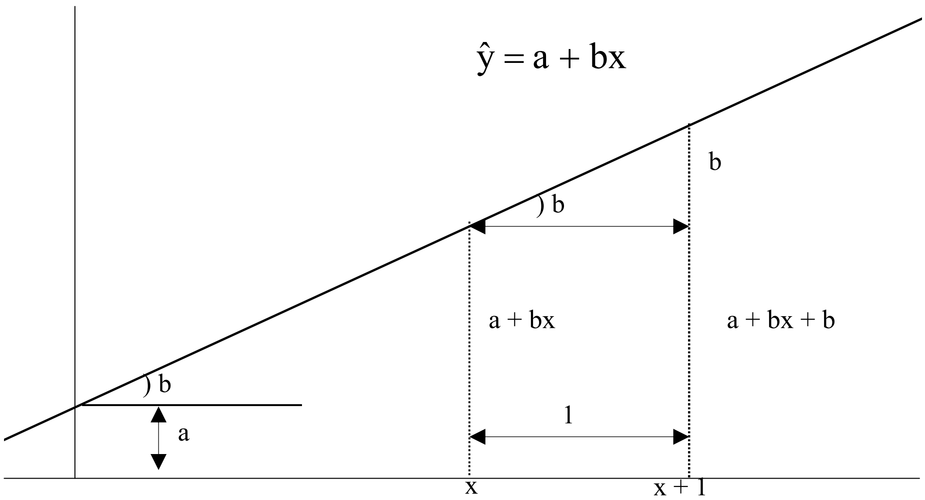


Figura 4. Representação do modelo $\hat{y} = a + bx$

No exemplo: $\hat{y}_i = -0{,}98 + 2{,}16x_i$, para x = 14, $\hat{y}$ = 29,26 e para x = 15, $\hat{y}$ = 31,42. A diferença entre os valores de $\hat{y}$ é 2,16, exatamente o valor de b; ou seja, para cada acréscimo de uma unidade em X, $\hat{y}$ acresce de 2,16. O intercepto a = –0,98 representa a quantidade de procaina hidrolisada para o tempo zero, o qual, neste caso, não possui significado biológico.

Observações:

(1) A regressão de y em x, $E(y/x) = -0{,}98 + 2{,}16.x_i$, representa, no caso do exemplo, a reta de regressão da quantidade de procaina hidrolisada sobre o tempo. Ou seja, E(y/x) nada mais é do que a média da distribuição de todas as quantidades de procaina hidrolisada em um dado tempo (x).

(2) O estimador de mínimos quadrados da variância de y dado x (σ^2), referido como quadrado médio residual, é dado pela fórmula:

$$s^2 = \hat{\sigma}^2 = \frac{\sum (y_i - \bar{y})^2 - \dfrac{[\sum (x_i - \bar{x})(\sum (y_i - \bar{y})]^2}{\sum (x_i - \bar{x})^2}}{n-2},$$

cuja estimativa, no exemplo, é 0,82. O que está se supondo é que esse valor é constante para cada x fixado (propriedade homoscedástica)

(3) Há situações nas quais X também aparece como uma variável aleatória. Nesses casos, pode ser que estejamos também interessados na regressão de
X em Y. Têm-se:

$$\hat{x}_i = \bar{x} \pm b_{X.Y}(y_i - \bar{y}), \text{ onde} \qquad b_{X.Y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2}$$

Tabela 2. Exemplo de regressão linear em planta entre área foliar (Y) e comprimento vezes a largura (X) de 20 folhas de bromélia selecionadas ao acaso:

X	0,08	0,15	0,08	0,05	0,08	0,11	0,08	0,10	0,06	0,05
Y	0,07	0,12	0,06	0,04	0,06	0,09	0,06	0,08	0,05	0,04

X	0,06	0,03	0,16	0,09	0,05	0,08	0,11	0,14	0,09
Y	0,05	0,03	0,13	0,07	0,03	0,06	0,09	0,11	0,08

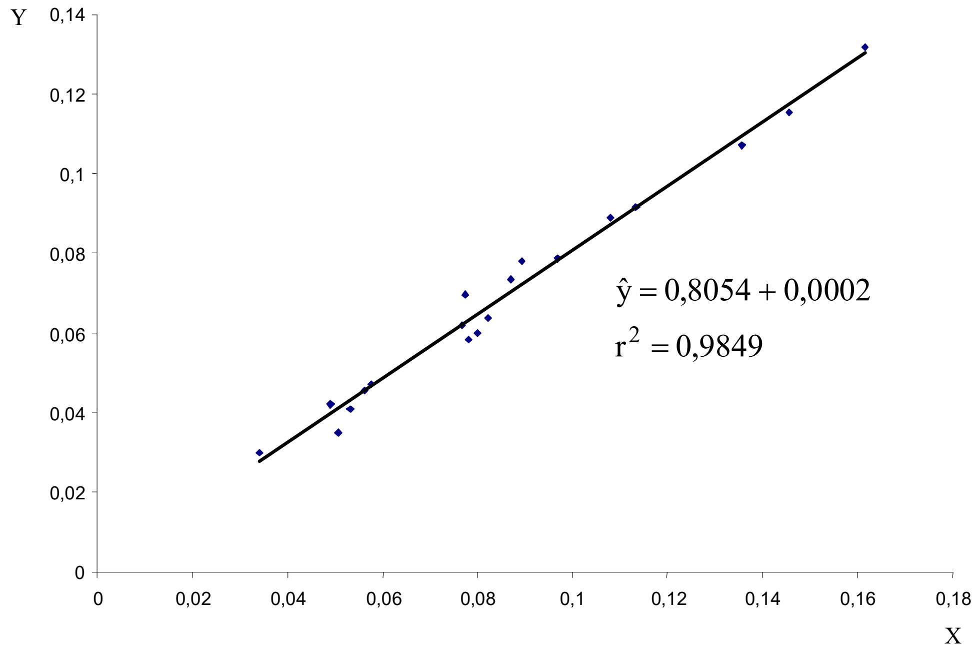


Figura 5. Área foliar (Y) em função do comprimento vezes a largura (X) da folha de bromélia.

12.4 Correlação

Vimos que numa análise de regressão linear simples, se determina, por meio de estimativas dos parâmetros, como uma variável X exerce, ou parece exercer efeito sobre uma outra variável Y.
Quando X e Y são ambas variáveis aleatórias, pode ser útil o conhecimento de uma medida que relacione as duas variáveis quando elas mantêm entre si uma relação dada por uma linha reta. Tal medida é dada pelo coeficiente de correlação (ρ). Assim, correlação é definida como a quantificação do
grau em que duas variáveis aleatórias estão relacionadas, desde que a relação seja linear.
Na análise de correlação se procura, então, determinar o grau de relacionamento entre as duas variáveis, ou seja, se procura medir a covariabilidade entre elas.

Na análise de regressão é necessário distinguir a variável dependente e a variável independente; na de correlação, tal distinção não é necessária.
No que segue, os dados são supostos **normalmente distribuídos**.

Definição: Sejam $x_1, x_2, ..., x_n; y_1, y_2, ..., y_n$ os valores observados de X e Y, respectivamente. Chama-se **coeficiente de correlação** (amostral**)** entre X e Y, o número dado por:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X).Var(Y)}} = \frac{\sum(y_i-\bar{y})(x_i-\bar{x})/n-1}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}.\frac{\sum(y_i-\bar{y})^2}{n-1}}} = \frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$$

Uma fórmula equivalente de cálculo de r, de fácil manuseio, é:

$$r = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sqrt{[\sum x_i^2 - (\sum x_i)^2/n][\sum y_i^2 - (\sum y_i)^2/n]}} = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

Propriedades

(1) O número r varia entre -1 e + 1

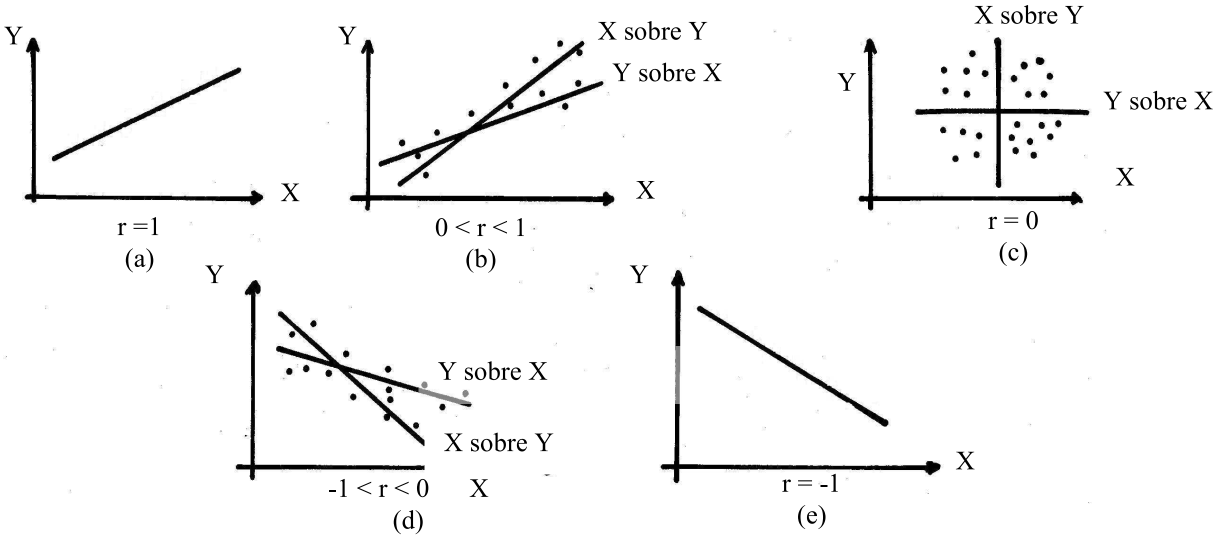


Figura 6. Retas de regressão e o coeficiente de correlação linear.

O valor numérico de r mede a intensidade da relação linear e o sinal de r indica o sentido da relação. Nas Figuras (a) e (e) há correlação perfeita: o valor de Y é determinado exatamente por uma reta linear em X, ou seja, os pontos estão dispostos de forma tal, que as retas de regressão de Y sobre X e de X sobre Y coincidem. Em (c), caso em que r = 0, o qual é interpretado como ausência de relação linear, os dois coeficientes de regressão $b_{Y.X}$ (Y em X) e $b_{X.Y}$ (X em Y) são também zero e, portanto, as retas de regressão são perpendiculares.
É importante assinalar que r = 0 não implica em ausência de relação entre duas variáveis. Isto é mostrado na Figura 7, onde apesar de r = 0, é evidente que existe uma relação parabólica entre X e Y. Portanto, r = 0 somente implica ausência de relação linear entre as duas variáveis.

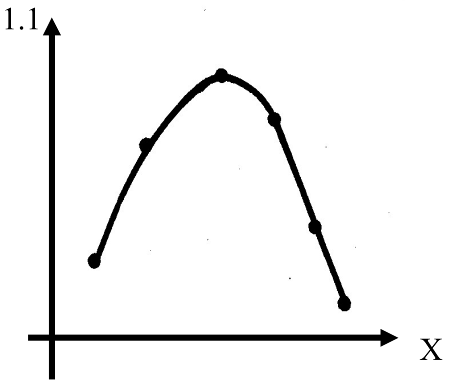


Figura 7. Relação parabólica entre X e Y, onde: r = 0.

(2) **r^2** é igual ao **coeficiente de determinação** da regressão linear simples ($\hat{y}_i = a + bx_i$). Note que $0 \leq r^2 \leq 1$.

O coeficiente de determinação pode ser interpretado como a proporção da variabilidade total observada entre os valores de Y, explicada pela regressão linear de Y sobre X ou seja,

$$r^2 = \frac{s_Y^2 - s_{Y/X}^2}{s_Y^2}$$

onde: $s_{Y/X}^2 = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$ é a variação dos valores de Y que ainda permanece, depois de se levar em conta a relação linear entre Y e X (devido ao fato que nem todos os pontos estão sobre a reta de regressão), que é parte não explicada pela regressão; e $(s_Y^2 - s_{Y/X}^2)$ é a variação em Y explicada pela regressão. Note que $s_{Y/X}^2$ envolve a soma dos desvios elevados ao quadrado das observações reais (y_i) dos valores ajustados ($\hat{y}_i$), isto é, $\sum_{i=1}^{n} e_i^2$, a qual é a quantidade minimizada ao se ajustar a linha de mínimos quadrados (veja Figura 8).

O coeficiente de determinação é, portanto, uma medida descritiva da qualidade do ajustamento obtido pela equação de regressão estimada. É particularmente importante quando é usado para fazer previsões e será tanto mais útil quanto mais próximo de um (1,0) estiver o seu valor. Se $r^2 = 1$, todos os dados na amostra situam-se na linha de mínimos quadrados; se $r^2 = 0$, não há uma relação linear entre X e Y.

Para o exemplo apresentado na Tabela 1, pode-se mostrar que $r^2 = (0{,}997)^2 = 0{,}994$. Esse valor implica em uma relação linear forte entre o tempo e a quantidade de procaina hidrolisada; em particular 99,4 % da variabilidade entre os valores observados de procaina hidrolisada é explicada pela relação linear entre essa variável e o tempo. O restante 1 – 0,994 = 0,006 (0,6 %) da variação não é explicada por essa relação.

(3) Das fórmulas do coeficiente de regressão e de correlação têm-se:

$$b_{Y.X} = r\frac{s_Y}{s_X} \qquad b_{X.Y} = r\frac{s_X}{s_Y}$$

onde: s_X e s_Y são os desvios padrão de X e Y, respectivamente.

Retas de regressão e o coeficiente de correlação linear

A equação da reta $\hat{Y} = a_1 + b_1 X$ ou a reta de regressão de Y em X, como visto, pode ser escrita sob a forma:

$$\hat{Y} = \overline{Y} + b_1(X - \overline{X}) \;\; ou \;\; \hat{Y} - \overline{Y} = b_1(X - \overline{X})$$

Como $b_1 = b_{Y.X} = r\dfrac{s_Y}{s_X}$

$$\hat{Y} - \overline{Y} = r\frac{s_Y}{s_X}(X - \overline{X}) \;\; ou \;\; y = r\frac{s_Y}{s_X} \cdot x \quad (1)$$

De modo semelhante, a reta de regressão de X em Y, $\hat{X} = a_2 + b_2 Y$, pode ser escrita como:

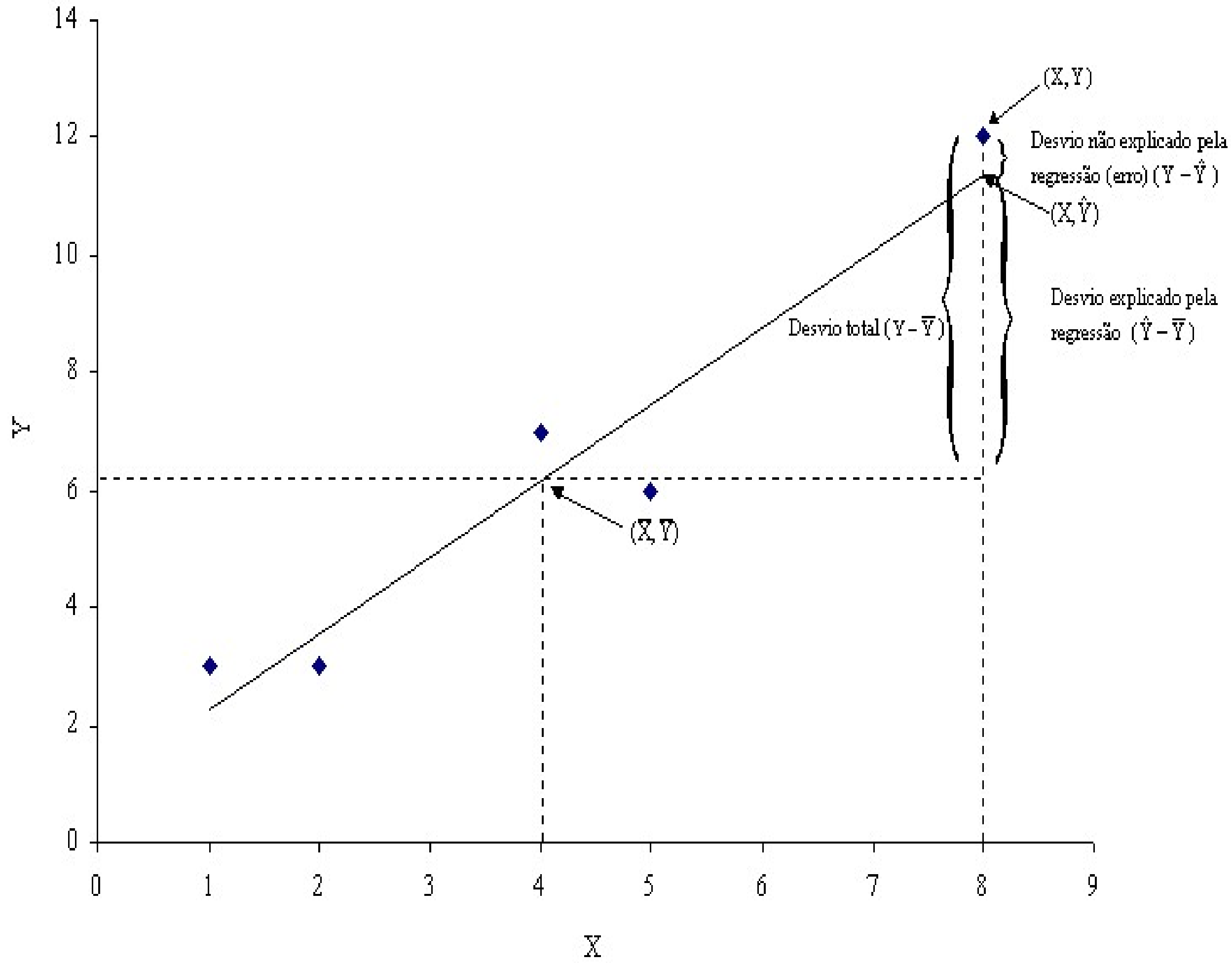


Figura 8. Regressão linear de Y sobre X

$$\hat{X}-\overline{X}=r\frac{s_X}{s_Y}(Y-\overline{Y}),\ \ onde\ b_2=b_{X.Y}=r\frac{s_X}{s_Y},\ \ ou\ \ \ x=r\frac{s_X}{s_Y}\cdot y\ \ \ (2)$$

As declividades das retas (1) e (2) somente serão iguais quando r = ± 1. Neste caso, as duas retas serão idênticas e há correlação linear perfeita entre as variáveis X e Y [Se r = ± 1, a equação (2) pode ser obtida da de (1) ou seja, $x=\frac{y}{s_Y/s_X}\,ou\ x=\frac{s_X}{s_Y}\cdot y$]. Quando r = 0, as

retas de regressão
estão em ângulo reto e não há correlação linear entre X e Y. Tais fatos estão ilustrados na Figura 6. Dessa forma, o coeficiente de correlação linear mede o afastamento angular entre as duas retas de regressão.

Note que: $b_1\cdot b_2=r\frac{s_Y}{s_X}\cdot r\frac{s_X}{s_Y}=r^2$, onde: r^2 = coeficiente de determinação.

12.5 Correlação e causa

É importante salientar que o coeficiente de correlação define apenas o sentido da variação conjunta das variáveis. A observação que duas variáveis tendem variar simultaneamente em uma direção ou em direções contrárias, onde os dados provavelmente indicariam uma correlação, positiva ou negativa, alta, não implicaria necessariamente na presença de uma relação de causa e efeito entre elas. Assim, na Figura 9, nota-se que existe uma correlação negativa entre o consumo de proteínas e o coeficiente de natalidade. Entretanto, isto não implica em afirmar que um aumento no consumo de proteínas determina redução da fertilidade. Portanto, uma correlação observada pode ser falsa (**correlação espúria**), isto é, pode ser devido a uma terceira e desconhecida variável causal.

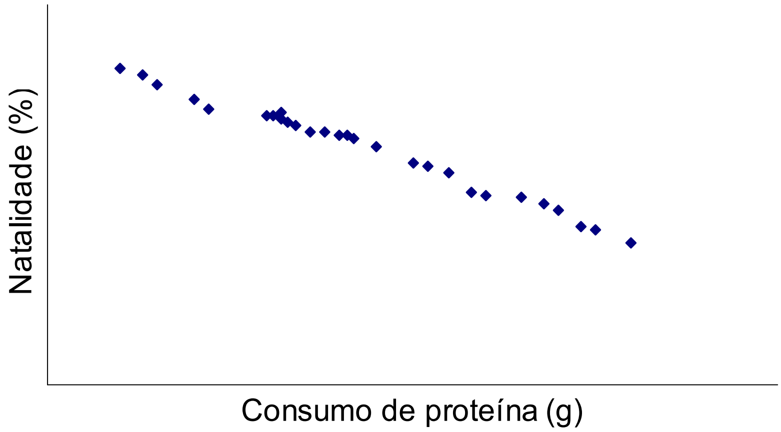


Figura 9. Diagrama de dispersão para o consumo individual diário de proteínas de origem animal e a natalidade, em 28 países.

Exemplo de correlação

Tabela 2. Amostra de pares de valores referentes aos pesos (kg) ao nascer (X) e aos 12 meses (Y) de 10 animais da raça Nelore:

X	29	32	28	23	28	34	27	24	27	20
Y	219	262	202	138	190	215	188	164	185	150

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}} = \frac{53202 - \frac{272.1913}{10}}{\sqrt{(7552 - 10.27,2^2)(377.743 - 10.191,3^2)}}$$

r = 0,87

Portanto, o grau de associação linear entre X e Y está quantificado em 87%.

12.6 4. Testes sobre o coeficiente de regressão (β) e correlação (ρ)

Verificaremos agora se os valores estimados de b e de r tem significados estatísticos, ou seja, se os coeficientes de regressão (β) e de correlação (ρ) que eles estão indicando são estatisticamente válidos.

A hipótese H_0: β = 0 (não existe dependência linear entre X e Y) pode ser testada usando a estatística:

$$\frac{b - \beta}{\sqrt{Var\,(b)}} = \frac{b}{\sqrt{Var(b)}},$$

que tem distribuição t com n – 2 graus de liberdade,

$$\text{onde:} \quad \text{var}\,(b) = \frac{\dfrac{\sum (y_i - \bar{y})^2 - \dfrac{[\sum (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum (x_i - \bar{x})^2}}{n-2}}{\sum (x_i - \bar{x})^2}.$$

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

Exemplo. Testar $H_0 : \beta = 0$ contra $H_1 : \beta \neq 0$ (existe dependência linear entre X e Y) empregando os dados apresentados na Tabela 1.
Solução:
n = 8 b = 2,16

$$\sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = 3.299{,}42 - \frac{(141{,}20)^2}{8} = 807{,}24$$

$$Var(b) = \frac{\dfrac{807{,}24 - \dfrac{(371{,}35)^2}{171{,}88}}{6}}{171{,}88} = \frac{0{,}82}{171{,}88} = 0{,}0048$$

$$t_{obs} = \frac{b - \beta}{\sqrt{Var(b)}} = \frac{2{,}16 - 0}{\sqrt{0{,}0048}} = 31{,}30$$

$\alpha = 5\%$ gl = n - 2 = 6 $t_{c\,(0{,}05;\,6)} = 2{,}447$

RC = {t > 2,447 ou t < -2,447}

Conclusão: como $t_{obs} \in$ a RC, rejeita-se H_0, com nível de significância de 5%. Sendo b = +2,16, há evidência de que os valores de Y realmente crescem com os valores de X.

Para testar $H_0 : \rho = 0$ (não existe correlação entre X e Y) contra $H_1 : \rho \neq 0$ (existe correlação entre X e Y) pode-se usar a estatística:

$$\frac{r - \rho}{\sqrt{Var(r)}}$$

que, para amostras retiradas de uma população para a qual $\rho = 0$, segue uma distribuição t com n – 2 graus de liberdade, onde: $Var(r) = \dfrac{1 - r^2}{n-2}$. Assim, $t = \dfrac{r\sqrt{n-2}}{\sqrt{1 - r^2}}$

Exemplo. Dos dados da Tabela 2,

$$t_{obs} = \frac{0{,}87\sqrt{10-2}}{\sqrt{1-(0{,}87)^2}} = \frac{2{,}46}{0{,}49} = 5{,}02$$

Se $\alpha = 0{,}01$, $t_{c\,(0{,}01;\,8)} = 3{,}355$.

Como $t_{obs} > t_c$, a hipótese nula é rejeitada ao nível de significância de 1%. Portanto, há evidência de que as variáveis X e Y são correlacionadas.

Obs.: pode-se mostrar que

$$\frac{b}{\sqrt{\mathrm{var}(b)}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Assim, para se testar a hipótese $\beta = 0$, pode-se usar a estatística $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \approx t(n-2)$, que é de cálculo mais fácil. No exemplo apresentado na Tabela 1,

$$\frac{b}{\sqrt{\mathrm{var}(b)}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}997\sqrt{8-2}}{\sqrt{1-(0{,}997)^2}} = 31{,}30$$

13 ANÁLISE BIDIMENSIONAL

13.1 Introdução

O objetivo da análise bidimensional é analisar o comportamento conjunto de duas variáveis. Assim, como para o caso de uma variável, a distribuição conjunta das frequências é um poderoso instrumento para ajudar a compreensão dos dados. A distribuição por frequência é representada por uma tabela de dupla entrada.

Exemplo 1. Usando as variáveis qualitativas avaliação ao nascer e sexo em bovinos, apresentadas no capítulo Estatística descritiva (Tabela 1), têm-se:

Tabela 1. Distribuição conjunta das frequências das variáveis avaliação ao nascer (X) e sexo (Y) dos animais da Fazenda Z.

Y	X			Total
	R	M	E	
Macho	3	12	5	20
Fêmea	7	18	5	30
Total	10	30	10	50

Fonte: Tabela 1 (Estatística descritiva)

A linha dos totais fornece a distribuição da variável X, e a coluna dos totais a distribuição da variável Y. Essas distribuições são chamadas de distribuições marginais, enquanto que a Tabela 1, constitui a distribuição conjunta de X e Y.
Em vez de se trabalhar com as frequências absolutas, pode-se construir tabelas com as frequências relativas (proporções), como foi feito no caso unidimensional. Mas aqui existem 3 possibilidades de se expressar a proporção de cada célula: em relação ao total geral (Tabela 2), ao total de cada linha e ao total de cada coluna (Tabela 3). De acordo com o resultado de cada pesquisa, uma delas será a mais conveniente de ser usada.

Tabela 2. Distribuição conjunta das proporções (em %), em relação ao total geral, das variáveis X e Y

Y	X			Total
	R	M	E	
Macho	6	24	10	40
Fêmea	14	36	10	60
Total	20	60	20	100

Fonte: Tabela 1

Os totais das margens fornecem as distribuições unidimensionais de cada uma das variáveis.

Tabela 3. Distribuição conjunta das proporções (em %), em relação aos totais de cada coluna, das variáveis X e Y

Y	X			Total
	R	M	E	
Macho	30,0 (15,0)	40,0 (60,0)	50,0 (25,0)	40,0 (100)
Fêmea	70,0 (23,3)	60,0 (60,0)	50,0 (16,7)	60,0 (100)
Total	100,0 (20,0)	100,0 (60,0)	100,0 (20,0)	100

Fonte: Tabela 1

Este tipo de distribuição serve para comparar a distribuição do sexo (Y) dos animais, conforme os níveis de avaliação ao nascer (X).
De modo análogo, pode-se construir a distribuição das proporções em relação ao total de linhas (valores entre parênteses na Tabela 3).

13.2 Independência de variáveis

Um dos principais objetivos de uma distribuição conjunta é descrever a associabilidade entre as variáveis, isto é, deseja-se conhecer o grau de dependência entre elas, de modo que se possa prever melhor o resultado de uma delas, quando se conhece o resultado da outra.
Vejamos, agora, como identificar a dependência ou não entre variáveis, por meio da distribuição conjunta, no caso entre X e Y (Tabela 1).
Inicialmente, deve-se construir as proporções segundo as linhas ou as colunas, para se fazer as comparações, pois fica difícil tirar alguma conclusão, devido à diferença entre os totais marginais. Fixando os totais das colunas, a distribuição está na Tabela 3. A partir dessa tabela pode-se observar (na coluna do total) que independentemente da avaliação, 40% dos animais são machos e 60% fêmeas. Havendo independência entre as variáveis, seria esperado estas mesmas proporções para cada nível de avaliação (R, M e E). Deste modo, a análise da Tabela 3 parece indicar haver **independência** entre as duas variáveis. Convém observar que a conclusão será a mesma, se for utilizado as proporções calculadas, mantendo-se constante os totais das linhas.
Por outro lado, se ao compararmos a distribuição das proporções pelos sexos, independentemente da avaliação (coluna de total), com as distribuições diferenciadas por nível de avaliação (colunas de R, M e E), observássemos uma disparidade bem acentuada nas proporções, então, neste caso, os resultados indicariam **dependência** entre as variáveis.

Exemplo 2. Vamos supor uma pesquisa envolvendo peso e sexo de bovinos, cuja distribuição conjunta é:

Tabela 4. Distribuição conjunta das frequências e proporções (em %), segundo o sexo (X) e o peso aos 12 meses de idade (Y), em kg

Y	X: Macho	Fêmea	Total
> 198	14 (70%)	8 (27%)	22 (44%)
≤ 198	6 (30%)	22 (73%)	28 (56%)
Total	20 (100%)	30 (100%)	50 (100%)

Fonte: Tabela 1 (Estatística descritiva)

Os totais entre parênteses indicam as proporções em relação aos totais das colunas.

Comparando-se a distribuição das proporções dos pesos, independentemente do sexo (coluna do total), com as distribuições diferenciadas por sexo (colunas de macho e fêmea), observa-se uma disparidade bem acentuada nas proporções, Assim, parece haver uma maior concentração de machos na classe de peso > 198kg e de fêmeas na classe ≤ 198kg. Portanto, nesse caso, as variáveis sexo e peso parecem **dependentes**.
Quando existe dependência entre variáveis é interessante quantificá-la.

2. Medida de dependência entre duas variáveis

De um modo geral, a quantificação do grau de dependência entre duas variáveis é feita pelos chamados coeficientes de associação ou correlação. Estes são medidas que descrevem num único número a dependência entre as duas variáveis. Para maior facilidade de compreensão, esses coeficientes usualmente variam de zero a um (ou, às vezes, de –1 até 1), e a proximidade de zero indica total independência.
A análise da Tabela 4 (Exemplo 2), mostra a existência de uma certa **dependência** entre as variáveis. Sob a **hipótese de independência**, os números esperados dentro de cada sexo são apresentados na Tabela 5.

Tabela 5. Valores esperados assumindo independência entre as variáveis X e Y

Y	X: Macho		Fêmea		Total
	% esp.	f_e	% esp.	f_e	
> 198	44	8,8	44	13,2	22 (44%)
≤ 198	56	11,2	56	16,8	28 (56%)
Total	100	20,0	100	30,0	50 (100%)

Fonte: Tabela 4 f_e = frequência esperada

Comparando as Tabelas 4 e 5, pode-se verificar as discrepâncias existentes entre os valores observados (Tabela 4) e os esperados (Tabela 5), assumindo independência entre as variáveis. Na Tabela 6 estão resumidos os desvios: observados menos esperados.

Tabela 6. Desvios entre os valores observados e esperados

Y	X						Total
	Macho			Fêmea			
	f_o	f_e	f_o-f_e	f_o	f_e	f_o-f_e	
> 198	14	8,8	5,2 (3,1)	8	13,2	-5,2 (2,0)	22 (44%)
≤ 198	6	11,2	-5,2 (2,4)	22	16,8	5,2 (1,6)	28 (56%)
Total	20	20,0	0	30	30,0	0	50 (100%)

Fonte: Tabelas 4 e 5 f_o = frequência observada

Obs: a soma total dos desvios é nula

Analisando-se a Tabela 6, pode-se constatar que à medida que a frequência observada (f_o) se aproxima da frequência esperada (f_e), a hipótese de independência está sendo verificada, e em caso das frequências esperadas se afastarem das observadas, isto é indicativo que a hipótese de independência não se verifica, ou seja as variáveis apresentam um certo grau de dependência.
Uma medida do afastamento global pode ser dada pela soma dos desvios relativos:

$$\frac{(f_o - f_e)^2}{f_e}$$

cujos valores são indicados entre parênteses na Tabela 6, para todas as células.

Chama-se essa medida de **qui-quadrado** (χ^2) e no Exemplo 2 têm-se:

$\chi^2 = 3,1 + 2,4 + 2,0 + 1,6 = 9,1$

No caso geral, a expressão de χ^2 é dada por

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$, onde a somatória é estendida a todas as células.

$$\text{Quando} \begin{cases} \chi^2 \to 0, \text{a tendência é de independência, e} \\ \chi^2 >>> ... > 0, \text{ a tendência é de dependência} \end{cases}$$

Assim, quanto maior for o valor de χ^2, maior será o grau de associação existente entre as duas variáveis. Mas fica difícil, baseando-se no valor de χ^2, julgar se associação é alta ou não. Por isso, várias medidas tem sido propostas:

(a) Coeficiente de contingência de Cramér

$$Q_1 = \frac{\chi^2}{n(q-1)}, \qquad 0 \leq Q_1 \leq 1$$

Se $Q_1^* = \sqrt{Q_1}$, $-1 \leq Q_1^* \leq 1$

(b) Coeficiente de contingência de Pearson

$$Q_2 = \sqrt{\frac{\chi^2}{n+\chi^2}}, \qquad 0 \le Q_2 \le \sqrt{\frac{q-1}{q}}$$

onde: q = número de linhas ou colunas da tabela, o que for menor
n = tamanho da amostra

Em (b), mesmo quando existe uma associação perfeita, Q_2 pode não ser igual a um (1,0). Uma alteração possível é considerar:

$$Q_2^* = \frac{Q_2}{\sqrt{\frac{q-1}{q}}}$$

(c) Coeficiente phi de Pearson em tabela 2 x 2

$$\varphi = \frac{(n_{11}\cdot n_{22} - n_{12}\cdot n_{21})}{\sqrt{n_{1.}\cdot n_{2.}\cdot n_{.1}\cdot n_{.2}}}, \qquad -1 \le \varphi \le 1$$

onde: n_{ij} representa a frequência conjunta observada da casela (i, j), $n_{i.}$ e $n_{.j}$ são os totais da i-ésima linha e da j-ésima coluna da tabela, respectivamente, i, j = 1, 2.

Grandes valores dessas medidas são indicativos de uma forte associação entre as variáveis, mas uma interpretação está faltando para valores pequenos e intermediários. Isto porque um tamanho amostral (n) grande, tende produzir medidas pequenas, embora o valor do χ^2 possa ser significativo. Trataremos da significância do χ^2 no capítulo sobre a distribuição desta estatística (teste χ^2).

Retornando ao Exemplo 2:

$$Q_1 = \frac{9{,}1}{50(2-1)} = 0{,}18 \qquad Q_2 = \sqrt{\frac{9{,}1}{50+9{,}1}} = 0{,}39$$

$$Q_2^* = \frac{0{,}18}{\sqrt{\frac{1}{2}}} = 0{,}25 \qquad \varphi = \frac{(14\cdot 22 - 6\cdot 8)}{\sqrt{20\cdot 30\cdot 22\cdot 28}} = 0{,}42$$

são indicadores do grau de associação entre as variáveis peso aos 12 meses e sexo.

Quanto ao Exemplo 1, os indicadores do grau de associação entre as variáveis avaliação ao nascer e sexo são:

$\chi^2 = 0{,}27$ $\qquad$ $Q_1 = 0{,}00540$

$Q_2 = 0{,}00537$ $\qquad$ $Q_2^* = 0{,}00759$, que sugerem, como já foi observado, independência entre as variáveis.

Exemplo 3. Amostras de leite de 50 vacas em lactação foram submetidas a dois tipos de testes: California mastitis test (A) e teste de Whitesid (B), para detecção de mastite sub-clínica. Os resultados foram os seguintes:

Teste B	Teste A						Total
	+			–			
	f_o	% esp.*	f_e*	f_o	% esp.*	f_e*	
+	25	60	18	05	60	12	30 (60%)
–	05	40	12	15	40	08	20 (40%)
Total	30	100	30	20	100	20	50 (100%)

Fonte: dados hipotéticos   *assumindo independência entre os testes

Verificar o grau de associação entre os dois testes.

Solução:
Os indicadores do grau de associação entre os testes são:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} = 2{,}72 + 4{,}08 + 4{,}08 + 6{,}12 = 17{,}0$$

$$Q_1 = \frac{\chi^2}{n(q-1)} = \frac{17{,}0}{50(2-1)} = 0{,}34 \qquad Q_1^* = \sqrt{Q_1} = \sqrt{0{,}34} = 0{,}58$$

$$Q_2 = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{17{,}0}{50+17{,}0}} = 0{,}50 \qquad Q_2^* = \frac{Q_2}{\sqrt{\frac{q-1}{2}}} = \frac{0{,}50}{\sqrt{\frac{1}{2}}} = 0{,}71$$

$$\varphi = \frac{\left(n_{11} \cdot n_{22} - n_{12} \cdot n_{21}\right)}{\sqrt{n_{1.} \cdot n_{2.} \cdot n_{.1} \cdot n_{.2}}} = \frac{(25.15 - 5.5)}{\sqrt{30.20.30.20}} = \frac{350}{600} = 0{,}58\,,$$

os quais sugerem, em geral, que estão associados.
Quando as variáveis envolvidas são ambas do tipo quantitativo, pode-se usar o mesmo tipo de análise apresentada anteriormente. De modo análogo, a distribuição conjunta pode ser resumida em tabelas de dupla entrada, e por meio das distribuições marginais é possível estudar a dependência ou não das variáveis. Algumas vezes, para evitar um grande número de entradas, agrupa-se os dados marginais em intervalos de classe, de modo semelhante ao resumo feito no caso unidimensional. Mas, além desse critério de análise, as variáveis quantitativas são passíveis de procedimentos analíticos mais refinados, para se verificar a associação entre elas. Dentre eles, um bastante útil é o **gráfico de dispersão**, que nada mais é do que a representação de pares de valores num sistema cartesiano.

13.3 Diagrama de dispersão

Vejamos a ilustração por meio de um exemplo.

Exemplo 4. Na Tabela 7 é apresentado os dados referentes aos pesos ao nascer (X) e aos 12 meses (Y) de idade de 10 bovinos de uma fazenda.

Tabela 7. Pesos ao nascer e aos 12 meses de idade, em kg, de 10 bovinos da Fazenda Z

Animal	Peso ao nascer (X)	Peso aos 12 meses (Y)
1	29	219
2	32	262
3	28	202
4	23	138
5	28	190
6	34	215
7	27	188
8	24	164
9	27	185
10	20	150

Fonte: Tabela 1 (Estatística descritiva)

Na Figura 1 estão representados os pares de pesos (X,Y) mostrados na Tabela 7.

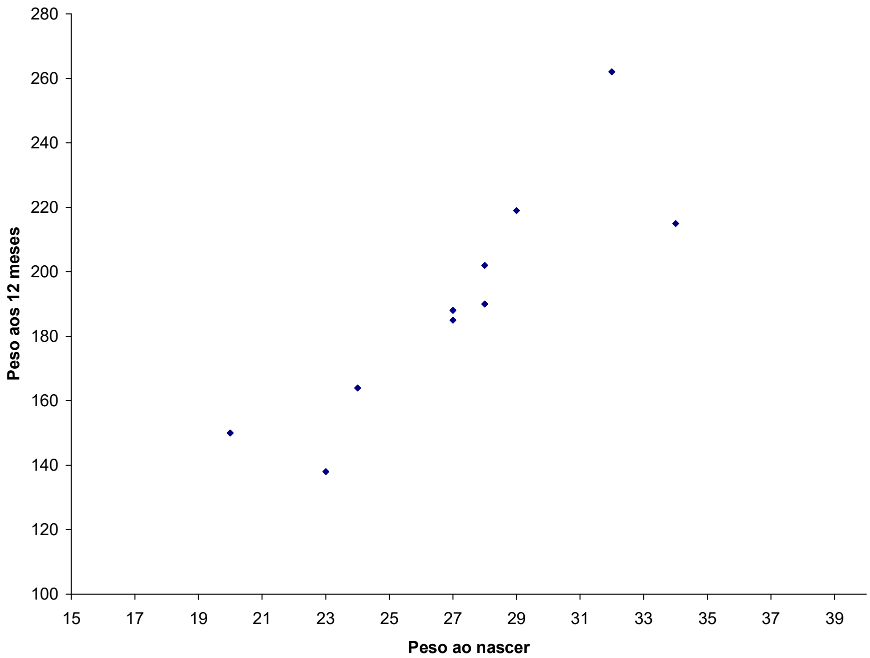


Figura 1. Diagrama de dispersão dos dados da Tabela 7

Por meio da observação da disposição dos pontos na Figura 1, conclui-se que parece haver uma dependência (positiva) entre as variáveis, porque no conjunto, à medida que o peso ao nascer aumenta, aumenta o peso aos 12 meses.

Outras possibilidades:

(a)

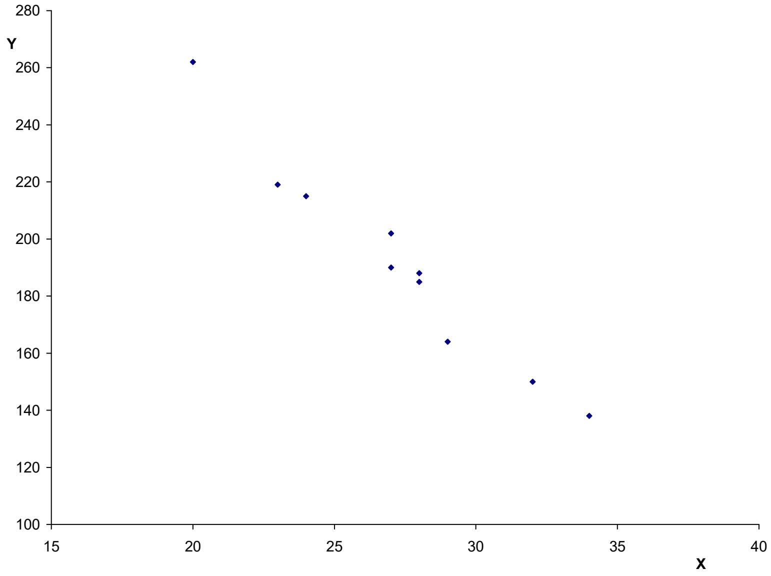


Figura 2. Diagrama de dispersão das variáveis X e Y

Observando-se o diagrama de dispersão da Figura 2, verifica-se que existe uma dependência inversa (ou negativa) entre as variáveis, isto é aumentando X, Y diminui.

(b)

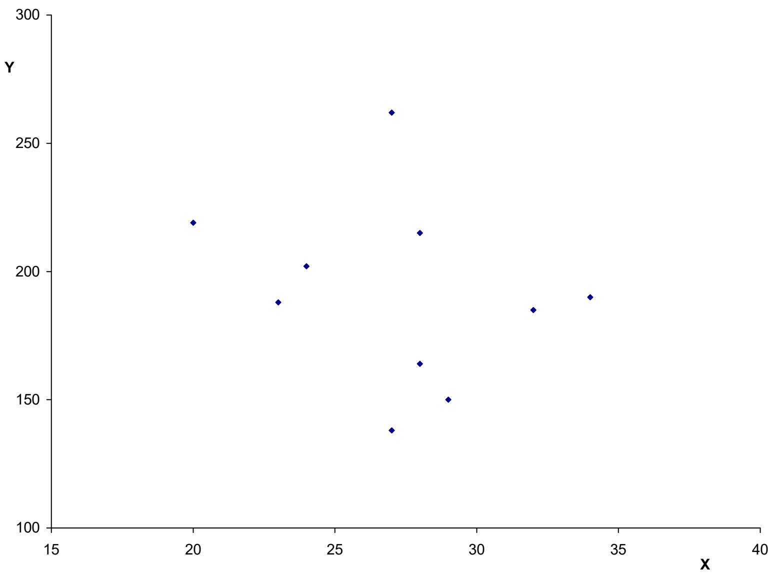


Figura 3. Diagrama de dispersão das variáveis X e Y

Do diagrama de dispersão da Figura 3, conclui-se que parece não haver dependência entre as duas variáveis.

A partir do apresentado, verifica-se que a representação gráfica das variáveis quantitativas ajuda muito a compreender o comportamento conjunto de duas variáveis quanto a existência ou não de associabilidade.

13.4 Coeficiente de correlação

Observada uma associação entre duas variáveis quantitativas é muito útil quantificar essa associabilidade. Existem muitos tipos de associação (linear, quadrática, cúbica) e o tipo de relação mais simples é a linear, onde é definido uma medida que julga o quanto a nuvem de pontos do diagrama de dispersão aproxima-se de uma reta. Essa medida é o **coeficiente de correlação**, que assume valores entre –1 e 1, e será tratada no último capítulo.

4. Distribuição hipergeométrica

Para se obter uma fórmula análoga àquela da distribuição binomial, aplicável a amostras **"sem reposição"**, caso em que os ensaios não são independentes, consideremos um conjunto de **N** elementos, dos quais **k** elementos são considerados sucessos e (**N – k**) como fracassos. Estaremos interessados, como na distribuição binomial, na probabilidade de se obter **x** sucessos em **n** ensaios, mas agora estaremos escolhendo, "sem reposição", **n** dos **N** elementos contidos no conjunto.

Note que há $\binom{k}{x}$ maneiras de escolher **x** sucessos dentre **k** possibilidades e $\binom{N-k}{n-x}$ maneiras de escolher (**n – x)** fracassos de (**N – k**) possibilidades e, portanto, $\binom{k}{x}\binom{N-k}{n-x}$ maneiras de escolher **x** sucessos e (**n – x**) fracassos (princípio fundamental da contagem). Por outro lado, desde que há $\binom{N}{n}$ maneiras de escolher **n** dos **N** elementos do conjunto, e assumindo que todas são igualmente prováveis (que é o que significa quando dizemos que a seleção é aleatória), segue-se que a probabilidade de **x** sucessos em **n** ensaios é:

$$(1) \quad P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \qquad \text{para x = 0,1, ..., n}$$

Assim, para amostras "sem reposição", a variável aleatória número de sucessos (**x**) em **n** ensaios, cuja função de probabilidade é dada por (1), é definida ter **distribuição hipergeométrica**, com parâmetros **n**, **N** e **k**. A média e a variância dessa distribuição são:

E(X) = np, onde: **p = k/N** (proporção populacional de sucessos), e

Var(X) = npq [(N – n)/N – 1)]

Quando **n/N** é pequeno, isto é, quando **n** é muito pequeno em relação a **N**, o fator **(N – n)/N – 1)** é próximo de 1, logo não há diferença prática entre extração sem e com reposição. Então, a distribuição hipergeométrica pode ser satisfatoriamente aproximada pela binomial, com **p = k/N** e **q = (N – k)/N**.

Comparando estas duas distribuições, podemos verificar que a binomial tem o mérito de simplicidade na fórmula de probabilidade. Ela tem como parâmetro a fração **p**, enquanto que a hipergeométrica requer o conhecimento de **k** e **N** individualmente.

Exemplo 1. Em problemas de controle de qualidade, lotes com **N** elementos são examinados. O número de elementos com defeito (k) é desconhecido. Colhe-se uma amostra de n elementos e determina-se o número de defeituosos na amostra (x). Como ilustração, suponha que, num lote de N = 100 vacinas, k = 10 estejam estragadas. Escolhendo-se n = 5 vacinas "sem reposição", calcular a probabilidade de não se obter vacinas estragadas (x = 0).

Solução:

$$P(X=x)=\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \qquad \Rightarrow \qquad P(X=0)=\frac{\binom{10}{0}\binom{90}{5}}{\binom{100}{5}}\cong 0{,}584$$

Usando a aproximação binomial:

$$p=P(E)=\frac{k}{N}=\frac{10}{100}=0{,}1 \qquad q=P(\overline{E})=0{,}9$$

$$P(X=0)=\binom{5}{0}0{,}1^0\,0{,}9^5=0{,}59$$

Exemplo 2. Suponha que em um lote com N = 20 animais existem k = 5 doentes. Escolhendo-se 4 animais do lote ao acaso, isto é, uma amostra de n = 4 elementos, de modo que a ordem dos elementos seja irrelevante, calcular a probabilidade de se obter x = 2 doentes na amostra.

Solução:

Usando (1):

$$\text{P(X = 2)} = \frac{\binom{5}{2}\binom{15}{2}}{\binom{20}{4}}=\frac{\frac{5!}{3!2!}\cdot\frac{15!}{13!2!}}{\frac{20!}{16!4!}}=\frac{10\,.\,105}{4.845}\cong 0{,}22$$

Sendo 4 doentes na amostra,

$$\text{P(X = 4)} = \frac{\binom{5}{4}\binom{15}{0}}{\binom{20}{4}}=\frac{5}{4.845}\cong 0{,}001$$

Usando a aproximação binomial:

$$p = P(D) = \frac{k}{N} = \frac{5}{20} = 0{,}25 \qquad q = P(\overline{D}) = 0{,}75$$

$$P(X = 2) = \binom{4}{2} 0{,}25^2\, 0{,}75^2 = 0{,}21$$

$$P(X = 4) = \binom{4}{4} 0{,}25^4\, 0{,}75^0 = 0{,}0039$$

14 VARIÁVEIS ALEATÓRIAS MULTIDIMENSIONAIS

14.1 Distribuição conjunta

Na maioria das vezes, ao se descrever os resultados de um experimento, se atribui a um mesmo ponto amostral os valores de duas ou mais variáveis aleatórias, indicando que os conceitos apresentados estendem-se facilmente ao conjunto formado de um número finito de variáveis aleatórias. Porém, o desenvolvimento será feito para variáveis aleatórias discretas.

Exemplo 1. Supondo que estamos interessados em estudar a composição de famílias de bovinos com 3 crias, quanto ao sexo, definamos:

X = número de machos

$$Y = \begin{cases} 1, \text{ se a primeira cria for macho} \\ 0, \text{ se a primeira cria for fêmea} \end{cases}$$

Z = número de vezes que houver variação do sexo entre um nascimento e outro, dentro de uma mesma família.

Com estas informações, e supondo que as possíveis composições tenham a mesma probabilidade, obtém-se a Tabela 1, onde, por exemplo, o evento MFM indica que a primeira cria é macho, a segunda é fêmea e a terceira é macho.

Tabela 1

Eventos	Prob.	X	Y	Z
MMM	1/8	3	1	0
MM F	1/8	2	1	1
M FM	1/8	2	1	2
F MM	1/8	2	0	1
M F F	1/8	1	1	1
F M F	1/8	1	0	2
F F M	1/8	1	0	1
F F F	1/8	0	0	0

Para cada uma das variáveis X, Y, Z, têm-se as respectivas distribuições de probabilidade. Por exemplo:

x	0	1	2	3
p(x)	1/8	3/8	3/8	1/8

y	0	1
p(y)	1/2	1/2

A Tabela 2 apresenta as probabilidades associadas aos pares de valores das variáveis aleatórias X e Y.

Tabela 2

(x,y)	(0,0)	(1,0)	(1,1)	(2,0)	(2,1)	(3,1)
p(x,y)	1/8	2/8	1/8	1/8	2/8	1/8

Nesta tabela, p(x, y) = p(X = x , Y = y) denota a probabilidade do evento (X = x e Y = y). A Tabela 2 é denominada distribuição conjunta de X e Y, que é uma distribuição bidimensional, isto e, de duas variáveis. Neste caso, uma maneira mais cômoda de representar a distribuição conjunta é através de uma tabela de duas entradas (Tabela 3).

Tabela 3

Y	X				p(y)
	0	1	2	3	
0	1/8	2/8	1/8	0	1/2
1	0	1/8	2/8	1/8	1/2
p(x)	1/8	3/8	3/8	1/8	1,0

14.2 Distribuições marginais

Da Tabela 3, pode-se obter facilmente as distribuições de X e Y. A primeira e a última colunas da tabela dão a distribuição de Y [y, p(y) = P(Y = y)], enquanto que a primeira e a última linhas da tabela dão a distribuição de X [x, p(x) = P(X = x)]. Estas distribuições são chamadas **distribuições marginais**.
Observa-se, pelo exemplo, que:

P(X = 1) = P(X = 1, Y = 0) + P(X = 1, Y = 1) = 2/8 + 1/8 = 3/8.

14.3 Variáveis aleatórias independentes

Exemplo 2. Consideremos agora a distribuição conjunta das variáveis Y e Z, definidas no exemplo 1. Da Tabela 1, obtém-se:

Tabela 4

Y	Z			p(y)
	0	1	2	
0	1/8	2/8	1/8	1/2
1	1/8	2/8	1/8	1/2
P(x)	1/4	2/4	1/4	1

Para essa tabela, observa-se que: $P(Z = z / Y = y) = \dfrac{P(Z = z, Y = y)}{P(Y = y)} = P(Z = z)$

para quaisquer z = 0, 1, 2 e y = 0, 1. O que mostra que

$P(Z = z, Y = y) = P(Z = z) \,.\, P(Y = y)$,

isto é, a probabilidade de cada casela é igual ao produto das respectivas probabilidades marginais. Por exemplo:

$P(Z = 1, Y = 1) = P(Z = 1) \,.\, P(Y = 1) = 2/4 \,.\, 1/2 = 1/4$

Também é verdade que $P(Y = y \,/\, Z = z) = P(Y = y)$ para todos os valores de Y e Z. Diz-se que Y e Z são **independentes**.

Definição. As variáveis X e Y, assumindo os valores $x_1, x_2, \ldots$ e $y_1, y_2, \ldots$, respectivamente, são independentes se, e somente se, para todo par de valores (x_i, y_i) de X e Y têm-se:

$P(X = x_i, Y = y_i) = P(X = x_i) \,.\, P(Y = y_i)$ (1)

Basta que (1) não se verifique para um par (x_i, y_i) para que X e Y não sejam independentes. Neste caso, diz-se que X e Y são **dependentes**.

Essa definição pode ser estendida para mais de duas variáveis aleatórias.

14.4 Funções de variáveis aleatórias

Retomemos a Tabela 3, que dá a distribuição conjunta das variáveis X e Y. A partir desta, pode-se considerar, por exemplo, a variável aleatória X + Y ou XY. A soma X + Y é definida naturalmente: a cada resultado do experimento, ela associa a soma dos valores de X e Y, isto é, (X + Y)(w) = (X)(w) + Y(w). Do mesmo modo, (XY)(w) = X(w).Y(w). Pode-se, então, construir a Tabela 5.

Tabela 5

(x_i, y_i)	$x + y$	xy	$p(x_i, y_i)$
(0, 0)	0	0	1/8
(0, 1)	1	0	0
(1, 0)	1	0	2/8
(1, 1)	2	1	1/8
(2, 0)	2	0	1/8
(2, 1)	3	2	2/8
(3, 0)	3	0	0
(3, 1)	4	3	1/8

A partir desta tabela, obtém-se as distribuições de X + Y e XY, ilustradas nas Tabelas 6 e 7.

Tabela 6

x + y	0	1	2	3	4
p(x+y)	1/8	2/8	2/8	2/8	1/8

Tabela 7

xy	0	1	2	3
p(xy)	4/8	1/8	2/8	1/8

Calculando as esperanças das variáveis X e Y da Tabela 3, obtém-se:

$$E(X) = \sum_{i=1}^{n} x_i\, p(x_i) = 0.1/8 + 1.3/8 + 2.3/8 + 3.1/8 = 12/8 = 3/2 = 1{,}5$$

$$E(Y) = \sum_{j=1}^{m} y_j p(y_j) = 0.1/2 + 1.1/2 = 1/2 = 0{,}5$$

Da Tabela 6, obtém-se:

$$E(X + Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i + y_j)\, p(x_i + y_j)$$

$$E(X + Y) = 0.1/8 + 1.2/8 + 2.2/8 + 3.2/8 + 4.1/8 = 16/8 = 2$$

Nota-se que E(X + Y) = E(X) + (Y).

Teorema 1. Se X é uma variável aleatória com valores $x_1, x_2, \ldots, x_n$ e probabilidades $p(x_1), p(x_2), \ldots, p(x_n)$, e Y é uma variável aleatória com valores $y_1, y_2, \ldots, y_m$ e probabilidades $p(y_1), p(y_2), \ldots, p(y_m)$ e se $p(x_i,y_i) = p(X = x_i, Y = y_i)$, i = 1, 2, ... , n ; j = 1, 2, ... , m, então:

E(X + Y) = E(X) + E(Y)

Isto é sempre verdade, quer seja X e Y independentes ou não.

Da Tabela 7, obtém-se:

$$E(XY) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i y_j \, p(x_i, y_j) = 0.4/8 + 1.1/8 + 2.2/8 + 3.1/8 = 8/8 = 1,0$$

Neste caso, observa-se que:

$E(XY) = 1,0 \neq E(X) \,.\, E(Y) = 1,5 \,.\, 0,5$, ou seja, de um modo geral, a esperança

de um produto não é o produto das esperanças. No entanto,

Teorema 2. Se X e Y são variáveis aleatórias independentes, então:

$$E(XY) = E(X) \,.\, E(Y) \qquad (2)$$

ou $$\sum_{i=1}^{n} \sum_{j=1}^{m} x_i y_j p(x_i, y_j) = \sum_{i=1}^{n} x_i p(x_i) \sum_{j=1}^{m} y_j p(y_j)$$

Isto pode ser mostrado a partir do exemplo 2 (Tabela 4), onde as variáveis aleatórias Y e Z são independentes. Sendo

$E(Z) = 0.1/4 + 1.2/4 + 2.1/4 = 2/4 + 2/4 = 1,0$
$E(Y) = 0.1/2 + 1.1/2 = 1/2$
$E(ZY) = 0.1/8 + 0.2/8 + 0.1/8 + 0.1/8 + 1.2/8 + 2.1/8 = 2/8 + 2/8 = 4/8 = 1/2$

então, $E(ZY) = E(Z) \,.\, E(Y)$

A recíproca do Teorema 2 não é verdadeira, ou seja, a expressão (2) pode ser válida e X e Y não serem independentes. Este fato é mostrado por meio de um exemplo.

Exemplo 3. Sejam X e Y variáveis aleatórias com a seguinte distribuição conjunta:

Tabela 8

	X			
Y	0	1	2	p(y)
1	3/20	3/20	2/20	8/20
2	1/20	1/20	2/20	4/20
3	4/20	1/20	3/20	8/20
p(x)	8/20	5/20	7/20	1,0

Observe que X e Y não são independentes, pois:

$P(X = 0, Y = 1) = 3/20 \neq P(X = 0) \,.\, P(Y = 1) = 8/20 \,.\, 8/20 = 4/25$. No entanto,

têm-se que:

$E(X) = 0 \,.\, 8/20 + 1 \,.\, 5/20 + 2 \,.\, 7/20 = 0{,}95$
$E(Y) = 1 \,.\, 8/20 + 2 \,.\, 4/20 + 3 \,.\, 8/20 = 2{,}00$
$E(XY) = 0 \,.\, 3/20 + 1 \,.\, 3/20 + 2 \,.\, 2/20 + 0 \,.\, 1/20 + 2 \,.\, 1/20 + 4 \,.\, 2/20 + +$
$0 \,.\, 4/20 + 3 \,.\, 1/20 + 6 \,.\, 3/20 = 38/20 = 1{,}9$

logo, $E(XY) = E(X) \,.\, E(Y)$
$1{,}9 = 2{,}0 \,.\, 0{,}95 = 1{,}9$

Obs.: (1) havendo um número finito de variáveis aleatórias $X_1, X_2, \dots, X_n$, então:
$E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$

(2) e se $X_1, X_2, \dots, X_n$ são variáveis aleatórias independentes, então:

$E(X_1 \,.\, X_2 \,.\, \dots \,.\, X_n) = E(X_1) \,.\, E(X_2) \,.\, \dots \,.\, E(X_n)$

14.5 Covariância de duas variáveis aleatórias

Será considerada agora uma medida numérica da variação conjunta de duas variáveis aleatórias.

Definição. Se X e Y são duas variáveis aleatórias, a covariância de X e Y é definida por

$$Cov(X,Y) = E\{[X - E(X)].[Y - E(Y)]\}\,, \qquad (3)$$

ou seja, o valor médio do produto dos desvios de X e Y em relação às suas respectivas médias.

Intuitivamente, pode-se dizer que X e Y variam na mesma direção se a probabilidade é alta que pequenos (grandes) valores de X estão associados com pequenos (grandes) valores de Y. Nesse caso, ambos os valores dos desvios $[X - E(X)]$ e $[Y - E(Y)]$ são positivos ou negativos com uma probabilidade alta, tal que o produto $[X - E(X)].[Y - E(Y)]$ é predominantemente positivo. Consequentemente, o valor esperado do produto é positivo e alto. Por outro lado, se X e Y tendem a variar em direções opostas, valores positivos de $[X - E(X)]$ estão mais frequentemente associados com valores negativos de $[Y - E(Y)]$ e vice-versa. O produto é então predominantemente negativo e o valor esperado é negativo. Neste sentido, o sinal e a magnitude de $[X - E(X)].[Y - E(Y)]$ refletem, respectivamente, a direção e a intensidade da relação linear entre X e Y, de modo que a covariância pode ser positiva ou negativa e teoricamente pode variar de $-\infty$ a $+\infty$.
Suponha que X assuma os valores $x_1, \dots, x_n$ e Y os valores $y_1, \dots, y_m$, e que $P(X = x_i, Y = y_j) = p(x_i, y_j)$. Então, Cov(X,Y) pode ser escrita como:

$$Cov(X,Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} [x_i - E(X)][y_j - E(Y)]\, p(x_i, y_j)$$

A primeira fórmula (3) pode ser escrita de uma forma mais simples:

$$Cov(X,Y) = E\,[XY - X \,.\, E(Y) - Y.\, E(X) + E(X) \,.\, E(Y)]$$
$$= E(XY) - E(X) \,.\, E(Y) - E(Y) \,.\, E(X) + E(X) \,.\, E(Y), \text{ ou seja,}$$

Cov(X,Y) = E(XY) – E(X) . E(Y)

Exemplo 4. Para as variáveis aleatórias X e Y da Tabela 3, obteve-se:

E(X) = 1,5 E(Y) = 0,5 e E(XY) = 1,0,

de modo que Cov(X,Y) = 1,0 – (1,5)(0,5) = 0,25

Definição. Quando Cov(X,Y) = 0, diz-se que X e Y são **não-correlacionadas**.

Exemplo 5. Considerando a distribuição conjunta de X e Y dada no exemplo 3 (Tabela 8):

E(X) = 0,95 E(Y) = 2,00 E(XY) = 1,90

logo, Cov(X,Y) = 1,90 – (0,95)(2,00) = 0

Exemplo 6. Retornemos à Tabela 4, onde foi verificado que as variáveis aleatórias Y e Z são independentes.

E(Z) = 1,0 E(Y) = 1/2 E(YZ) = E(Z) . E(Y) = 1/2
logo, Cov(Y, Z) = E(YZ) – E(Y) . E(Z) = 1/2 – 1 . 1/2 = 0

Proposição 1. Se X e X Y são duas variáveis aleatórias independentes, então E(X.Y) = E(X).E(Y) e Cov(X,Y) = 0.

De outro modo, se X e Y são independentes, isto implica X e Y não-correlacionadas. A recíproca não é verdadeira, isto é, Cov(X,Y) = 0 não implica X e Y independentes. De fato, para as variáveis aleatórias X e Y do exemplo 3 (Tabela 8), Cov(X,Y) = 0, mas como foi verificado, X e Y não são independentes.

Teorema 3. Para as duas variáveis aleatórias X e Y, escrevendo Z = X + Y, sempre temos $\mu_Z = \mu_X + \mu_Y$, e

$$\text{Var}(X+Y) = \text{Var}(Z) = E(Z-\mu_Z)^2 = E[(X-\mu_X)+(Y-\mu_Y)]^2$$

$$= E[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)]$$

$$= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$$

Do mesmo modo obtemos a variância da diferença de duas variáveis, isto é,

Var(X - Y) = Var(X) + Var(Y) - 2Cov(X,Y); e

(c) Se X e Y são independentes, então:

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$$

As relações (a) e (b) podem ser generalizadas para mais de duas variáveis aleatórias. Em particular, se $X_1, X_2, \ldots, X_n$ são variáveis aleatórias independentes, então

$$Var(X_1 \pm \ldots \pm X_n) = Var(X_1) + \ldots + Var(X_n)$$

A covariância isoladamente não é conveniente como uma medida da relação entre duas variáveis. Ela depende da unidade na qual X e Y são medidos. Se estivermos estudando a dependência entre as variáveis X: peso do pai em kg e Y: peso do filho em kg, ao calcularmos a covariância, teremos uma medida ao quadrado (kg^2). Além disso, o campo de variação é muito amplo, isto é,
$-\infty < Cov(X, Y) < +\infty$. Assim, como uma medida de relação linear que não depende de qualquer espécie de unidade, será considerado um índice chamado **coeficiente de correlação linear** ou simplesmente **coeficiente de correlação**.

Definição. O coeficiente de correlação de X e Y é definido por:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma(X).\, \sigma(Y)}$$

Exemplo 7. (a) Para as variáveis X e Y do exemplo 3 (Tabela 8), Cov (X, Y) = = 0. Portanto, $\rho(X, Y) = 0$.

(b) Para as variáveis X e Y do exemplo 1 (Tabela 3) têm-se:

Cov (X,Y) = 0,25 E(X) = 3/2

$$Var(X) = \sum_i [x_i - E(X)]^2 \cdot p(x_i)$$
$$= (0 - 3/2)^2.1/8 + (1 - 3/2)^2.3/8 + (2 - 3/2)^2.3/8 + (3 - 3/2)^2.1/8 = 0{,}75$$

E(Y) = 1/2

$$Var(Y) = \sum_j [y_j - E(Y)]^2 p(y_j) = (0 - 1/2)^2.1/2 + (1 - 1/2)^2.1/2 = 0{,}75$$

logo, $$\rho(X, Y) = \frac{0{,}25}{\sqrt{(0{,}75)(0{,}25)}} = 0{,}58$$

O coeficiente de correlação é uma quantidade adimensional e tem as seguintes propriedades:

i) $\rho(X, Y) = \rho(Y, X)$
ii) $-1 \leq \rho(X,Y) \leq 1$
iii) $\rho(X, X) = 1$ $\rho(X,-X) = -1$

Quando $\rho(X,Y) = \pm 1$, existe uma correlação perfeita entre X e Y, isto é, $Y = a + bX$; se $\rho(X,Y) = 1$, $b > 0$, e se $\rho(X,Y) = -1$, $b < 0$. O grau de associação linear entre X e Y varia à medida que $\rho(X,Y)$ varia entre –1 e +1.

15 BIBLIOGRAFIA

APOSTILA PRÁTICA DE ESTATÍSTICA BÁSICA - LAVRAS

BHATTACHARYYA, G.K.; JOHNSON, R.A. Statistical concepts and methods. New York: John Wiley & Sons, Inc., 1977.

BUSSAB, W.O.; MORETTIN, P.A. Estatística básica. São Paulo: Saraiva, 2003.

ELANDT-JOHNSON, R.C. Probability models and statistical methods in Genetics. New York: John Wiley & Sons, Inc., 1971.

MAGALHÃES, M.N.; LIMA, A.C.P. Noções de probabilidade e estatística. São Paulo: Edusp, 2002.

PETRIE, A.; WATSON, P. Estatística em ciência animal e veterinária. São Paulo: Editora Roca Ltda, 2009.

PORTAL ACTION: Ambiente Virtual de Aprendizado. URL http://www.portalaction.com.br/ambiente-virtual-de-aprendizado. Último acesso em 14 de março de 2016.

RAO, P.V. Statistical research methods in the life sciences. Pacific Grove: Brooks/Cole Publishing Company, 1998.

SIEGEL, S.; CASTELLAN JR, N.J. Nonparametric statistics for the behavioral sciences. 2.ed. New York: McGraw-Hill, 1988

SOARES, J.F.; FARIAS, A.A.; CESAR, C.C. Introdução à estatística. Rio de Janeiro: Guanabara Koogan S.A., 1991.

THOMPSON, S.K. Sampling. New York: John Wiley & Sons, Inc., 1992.

ZAR, J. H. Biostatistical analysis. New Jersey: Prentice Hall, 1999.

MEYER, P. L. Probabilidade Aplicações à Estatística. Rio de Janeiro: Ao Livro Técnico S.A, 1974.