Idéias em Estatística

Renato Vicente

GRIFE/EACH/USP 3/06/09

Sumário

Capítulo 1

Idéias Centrais em Estatística

1.1 Variáveis Aleatórias

A idéia mais central da Estatística é o fato de que medidas em geral têm precisão limitada (simples contagens têm, no entanto, precisão infinita) e observações apresentam variações não controladas. Dessa maneira, a mensuração repetida da mesma quantidade resultará em valores distintos. Variações que se devam a fatores que não controlamos constituem o que denominamos *erros estatísticos*. Assim, uma medida de altura sem controle da variável "sexo" produzirá resultados com variações estatísticas devido a diferenças na altura devido ao gênero. Para ilustrarmos uma variável aleatória recorremos ao histórico experimento de Cavendish para determinação da massa da Terra.

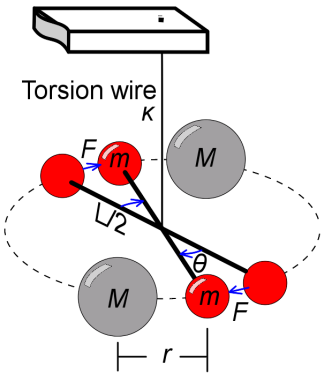


Figura 1.1: Balança de torção de Cavendish.

O experimento de Cavendish foi publicado em 1798 e consistia em uma haste de madeira de comprimento $L = 1.8m$ com esferas de aço com massas $m = 730g$ nas pontas suspensa por um fio. Em uma posição próxima a cada uma duas esferas pequenas eram colocadas duas esferas grandes com massas $M = 158kg$. Devido à atração gravitacional entre as esferas e o torque do fio as esferas pequenas tendem a uma posição de equilíbrio com os centros das esferas grande e pequena permanecendo a uma distância r após uma deflexão de um ângulo θ da situação inicial. Medindo o coeficiente de torção do fio (k) pela

observação de oscilações naturais da haste, a distância de equilíbrio r e o ângulo de deflexão θ. Cavendish conseguiu efetuar medidas da densidade da Terra em relação à densidade da água. O experimento é claramente extremamente sensível e as fontes possíveis de erros estatísticos são inúmeras. Cavendish realizou $n = 29$ repetições do experimento produzindo o seguinte rol de dados:

4,88	5,07	5,10	5,26	5,27	5,29	5,29	5,30	5,34	5,34
5,36	5,39	5,42	5,44	5,46	5,47	5,50	5,53	5,55	5,57
5,58	5,61	5,62	5,63	5,65	5,68	5,75	5,79	5,85	

Tabela 1.1: Rol dos dados para medidas da densidade da Terra relativos à densidade da água. O conjunto dos dados é uma *amostra* de tamanho $n = 29$.

Os valores apresentam variações que não são arbitrárias. Como alguns intervalos de valores ocorrem com mais freqüência do que outros podemos atribuir a eles um certo elemento de realidade. De certa forma, a constatação que dá origem a Estatística é a de que não é possível falarmos de um valor para uma medida, mas apenas de uma distribuição de valores.

Uma *variável aleatória* é um objeto matemático X que a cada medida fornece um valor x. Uma medida x tem probabilidade $p(x)\triangle x$ de ser encontrada em um intervalo de x a $x + \triangle x$. Dizemos que X está distribuída segundo $p(x)$ e indicamos isso como $X \sim p(x)$.

Há três formas básicas para visualizarmos a distribuição de uma variável aleatória.

1.1.1 Diagrama de Ramos e Folhas

O diagrama ramo e folhas é uma forma de organização dos dados que permite a utilização dos próprios dígitos para desenho da distribuição da variável aleatória. Para isso escolhemos o último digito como folhas e os restantes como ramos da seguinte maneira:

```
48 | 8
49 |
50 | 7
51 | 0
52 | 6799
53 | 04469
54 | 2467
55 | 03578
56 | 12358
57 | 59
58 | 5
```

Figura 1.2: Diagrama de ramos e folhas: cada dígito ocupa a mesma largura e a curva formada pelo envoltório representa a distribuição da variável aleatória que nos interessa, a saber, a densidade da Terra medida em termos da densidade da água.

Se tivéssemos que apostar em um valor representativo para a densidade da Terra, com base em nossa amostra, deveríamos escolher algo entre 5,20 e 5,70. Uma primeira estimativa seria, por exemplo, 5,45.

1.1.2 Tabela de Freqüências

A segunda alternativa consiste na construção de uma tabela de freqüências. Para isso escolhemos a largura dos intervalos de contagem de forma a termos a tabela adequadamente povoada (ou seja, sem muitos intervalos sem nenhuma contagem nem com poucos intervalos largos demais). No caso de nossos dados utilizamos o bom senso para escolhermos um intervalo de largura $0,2$. Temos então:

Intervalo	n_i	f_i (%)
$4,8 \vdash 5,0$	1	$\frac{1}{29} \times 100 = 3,5$
$5,0 \vdash 5,2$	2	7
$5,2 \vdash 5,4$	9	$\frac{9}{29} \times 100 = 31$
$5,4 \vdash 5,6$	9	31
$5,6 \vdash 5,8$	7	24
$5,8 \vdash 6,0$	1	$3,5$
Total	29	100

Tabela 1.2: O símbolo $\vdash$ significa que o lado esquerdo está incluído no intervalo e o direito excluído. A última coluna contém as freqüências relativas que indicam a chance de encontrarmos medidas de nossa amostra em cada intervalo.

Como a variável que medimos é quantitativa e contínua (ou seja, tem precisão limitada) somos obrigados a fixar intervalos para construirmos a tabela de freqüências. Ao escolhermos estes intervalos perdemos informação detalhada do rol, ou seja, não sabemos mais quais seriam as nove medidas no intervalo 5,4 a 5,6, por exemplo. Dizemos então que construímos uma tabela de freqüências com perda de informação.

1.1.3 Histograma

A representação gráfica por excelência para a distribuição de uma variável aleatória é o histograma.

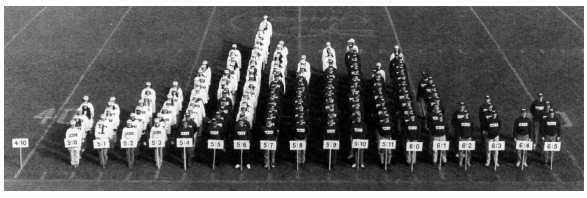


Figura 1.3: Histograma vivo para alturas de estudantes. Mulheres em branco, homens em preto. As alturas em pés (1 pé=30.5cm). Note que duas populações diferentes são aparentes. A separação em dois histogramas diminuiria a largura de cada um dos histogramas resultantes. O histograma sugere uma distribuição bimodal de alturas (uma moda para cada sexo).

Na versão mais simples apenas representamos a tabela de freqüências em um gráfico com barras da seguinte maneira.

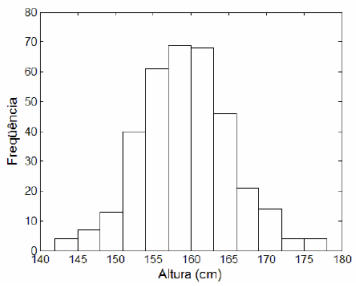


Figura 1.4: Os riscos no eixo horizontal indicam a localização das medidas. O histograma é uma representação da contagem dos pontos em cada intervalo e nos dá uma idéia imediata da distribuição da variável aleatória.

Pela definição de variável aleatória a distribuição $p(x)$ nos dará a chance de encontrarmos as medidas em um certo intervalo de largura $\triangle x$ quando calcularmos o produto $p(x)\triangle x$. No nosso exemplo $\triangle x = 0,2$. Se quisermos saber $p(x)$ precisamos que em cada intervalo i $0,2 \times p_i = f_i$, onde f_i são as freqüências relativas tabeladas acima. Assim, por exemplo, no intervalo $4,8 \vdash 5,0$ teremos que $p_1 = 0,035/0,2 = 0,175$, $p_2 = 0,07/0,2 = 0,35$,$p_3 = 0,31/0,2 = 1,55$ e assim por diante. A função $p(x)$ é denominada *densidade de probabilidade* e serve para conectarmos as observações empíricas de variáveis aleatórias à teoria matemática das probabilidades. A partir do histograma podemos então construir um gráfico para a densidade de probabilidade o que fazemos a seguir.

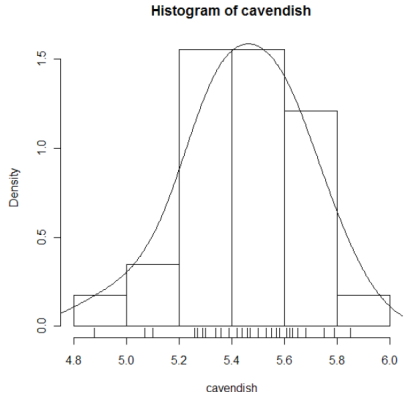


Figura 1.5: Densidade de probabilidade. A curva sobreposta é uma tentativa de estimar qual seria o aspecto geral do histograma se repetíssemos as medidas muitas vezes e construíssemos um novo histograma com divisões bem finas. A área de um intervalo abaixo da densidade corresponde à probabilidade de ocorrência deste intervalo de valores da variável aleatória.

O histograma obtido a partir de uma amostra nos traz informação sobre a densidade de probabilidade. De forma geral podemos dizer que:

> O histograma é uma *estimativa amostral* da densidade de probabilidade $p(x)$ que define a variável aleatória X.

1.2 Medidas Resumo

A questão que tentamos responder experimentalmente continua, no entanto, em aberto. Qual seria a densidade da Terra? Nossa amostra pode ser descrita por uma distribuição de valores que explicita as variações devido a erros estatísticos. Gostaríamos agora de extrair dessa amostra um valor que melhor representasse o que obteríamos se pudéssemos realizar as mesmas medidas livres de erros. Suponhamos que os erros aumentassem ou diminuíssem o valor "real" com probabilidades idênticas. O valor livre de erros deveria então estar no centro da distribuição observada. O centro da distribuição é representado por uma *medida de posição* ou *de tendência central*. O erro de nossas medidas é descrito pela largura da distribuição denominada *medida de dispersão ou variação*.

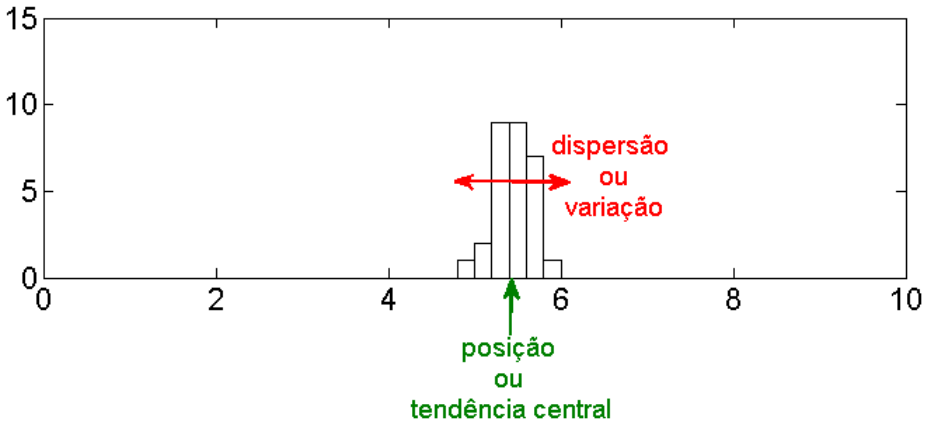


Figura 1.6: Uma distribuição pode ser descrita de forma resumida por uma medida de posição e uma medida de dispersão.

1.2.1 Medidas de Posição

Há três alternativas populares para medidas de posição: mediana, moda, média.

1.2.1.1 Mediana

A *mediana* é o menor valor maior do que 50% dos dados da amostra.

A determinação é muito simples: se o rol contiver um número ímpar de dados subtraímos um, dividimos por dois e pegamos o ponto do meio; se o número de dados for par, dividimos ao meio e utilizamos a média entre o último dado a esquerda e o primeiro à direita. Exemplificando em nosso rol teremos:

4,88	5,07	5,10	5,26	5,27	5,29	5,29	5,30	5,34	5,34
5,36	5,39	5,42	5,44	**5,46 ***	5,47	5,50	5,53	5,55	5,57
5,58	5,61	5,62	5,63	5,65	5,68	5,75	5,79	5,85	

Tabela 1.3: Temos $n = 29$. A mediana será então definida pelo o dado na posição (*rank ou ordem*) de número 15. Mediana$= 5,46$.

Note que alterações nos dados que não mudem a quantidade de dados em cada um dos lados da mediana não modificam seu valor. Por exemplo, se substituirmos $x_{29} = 5,85$ por $x_{29} = 585$ teremos o mesmo valor. Dizemos que a mediana é uma *medida de posição* ***robusta***.

1.2.1.2 Moda

A(s) *moda(s)* é(são) o(s) valor(es) mais freqüente(s).

A moda está definida também para variáveis nominais (por exemplo: nomes, marcas, localidades, sexo, estado civil). Além dessa aplicação óbvia, a moda é útil como ferramenta de diagnóstico para a presença de representantes de duas populações diferentes em uma amostra. Neste caso a moda será muito diferente da mediana e da média. Um exemplo ilustrativo é exibido a seguir:

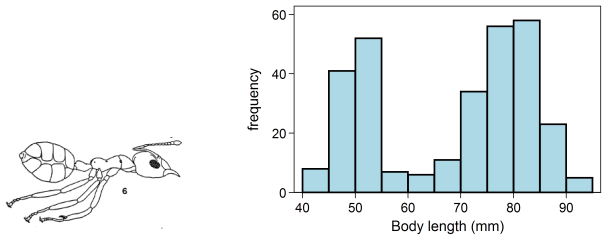


Figura 1.7: Histograma (à direita) para medidas do comprimento do corpo de formigas verdes operárias (à esquerda). O histograma pode ser considerado, desprezando-se pequenas flutuações estatísticas, como sendo bimodal. As modas seriam 52,5 mm e 82,5 mm. A bimodalidade sugere a presença de representantes de duas populações distintas na amostra. Note que a mediana estaria localizada entre as modas. A diferença entre modas e mediana é um indício da presença de duas populações distintas. Para decidirmos se há ou não duas populações seria necessário calcularmos o risco de estarmos errados caso decidíssemos que há duas populações. O caculo deste risco é parte do processo de *teste de hipóteses.*

Nos dados do experimento de Cavendish a moda pode ser localizada facilmente na tabela de freqüências ou no histograma como sendo MODA=5,4.

1.2.1.3 Média

A média aritmética pode ser interpretada como uma forma de utilizar o fato de erros estatísticos ocorrerem com igual probabilidade para cima e para baixo do valor "real" para obter estimativas mais precisas pela replicação de um experimento. A média em uma amostra é definida da seguinte forma:

$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{j=1}^{n} x_j$, onde x_1,x_2,...,x_n são observações contidas na amostra.

Para nosso exemplo teremos $\overline{x} = 5,448$.

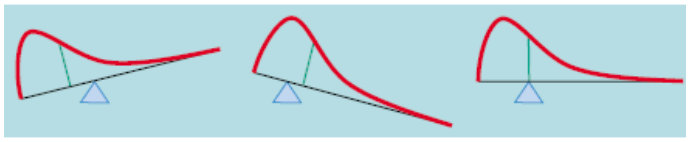


Figura 1.8: Se imaginarmos que cada dado é um pequeno peso posicionado na posição equivalente ao seu valor na reta, a média pode ser compreendida intuitivamente como o ponto de equilíbrio (centro de massa) do histograma. Dessa forma esperamos que uma pequena quantidade de valores extremos (como no lado direito da figura) domine o cálculo da média. A média é, portanto, uma medida pouco robusta, muito sensível a valores raros mas extremos.

1.2.2 Medidas de Dispersão

A cada medida de posição corresponde uma medida de dispersão, para as três medidas de posição que apresentamos acima temos, respectivamente: a distância interquartis, a largura à meia-altura e o desvio padrão.

1.2.2.1 Distância Interquartis

A distância interquartis (*inter-quartil ratio=IQR)* é

> A *distância interquartis* (*inter-quartil ratio=IQR)* é a largura da caixa que contém 50% dos dados no centro da amostra. $IQR = Q3 - Q1$, onde $Q1$ é o primeiro quartil definido pelo menor valor maior que 25% dos dados e $Q3$ é o terceiro quartil, definido como o menor valor maior que 75% dos dados.

Os quartis são obtidos ao calcularmos o sumário de cinco números que serve para construirmos o *boxplot*. O boxplot é uma forma alternativa para representação de uma distribuição, mais detalhada do que uma simples par de medidas resumo mas menos detalhado do que seria um histograma. Para o exemplo que estamos discutindo temos:

4,88	5,07	5,10	5,26	5,27	5,29	***5,29****	****5,30***	5,34	5,34
5,36	5,39	5,42	5,44	**5,46 ***	5,47	5,50	5,53	5,55	5,57
5,58	***5,61****	****5,62***	5,63	5,65	5,68	5,75	5,79	5,85	

Tabela 1.4: O primeiro quartil está localizado entre o sétimo e o oitavo dados sendo Q1=5,295. Já o terceiro quartil está situado entre o vigésimo segundo e o vigésimo terceiro dados Q3=5,615. IQR=5,615-5,295=0,32.

A distância interquartis pode ser visualizada ao compararmos o boxplot equivalente ao histograma dos dados.

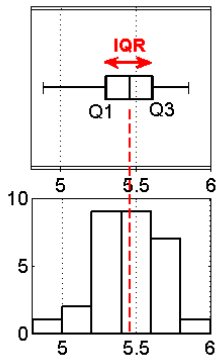


Figura 1.9: Box plot e histograma para o experimento de Cavendish. A linha tracejada representa a posição da mediana. Dentro da caixa temos 50% das observações. A largura da caixa é o IQR e representa a largura do histograma. No nosso exemplo IQR=0,32.

Como a mediana, o IQR também é uma *medida de dispersão* ***robusta***.

1.2.2.2 Largura a meia-altura

A *largura a meia-altura (LMA)* é a largura do histograma na metade da freqüência máxima.

A forma mais direta para determinação da LMA é através da tabela de freqüências ou do próprio histograma.

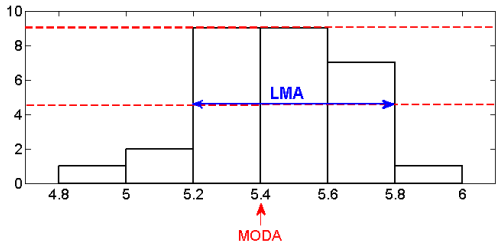


Figura 1.10: A LMA é determinada medindo-se a largura do histograma na metade da altura máxima. Em nosso exemplo LMA=0,6.

1.2.2.3 Desvio Padrão

O desvio padrão é uma medida da distância média dos dados em relação à média. Para determina-lo primeiro calculamos a variância.

VAR= $\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n}$ ou, para ficar com cara de livro de Estatística, VAR $=\frac{1}{n}\sum_{j=1}^{n}(x_j-\overline{x})^2$.

O desvio padrão é a raiz quadrada da variância:

$$\text{DP}=\sqrt{\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n}}=\sqrt{\text{VAR}}$$

Para ilustrar realizamos o calcula passo a passo em nosso exemplo. A média é $\overline{x}=5,45$ (aqui arredondamos para duas casas decimais para facilitar os cálculos). Começamos por calcular os resíduos $(x_j-\overline{x})$:

4,88-5,45=-0,57	-0,38	-0,35	-0,19	-0,18
-0.16	-0.16	-0.15	-0.11	-0.11
-0,09	-0,06	-0,03	-0,01	0,01
0,02	0,05	0,08	0,10	0,12
0,13	0,16	0,17	0,18	0,20
0,23	0,30	0,34	0,40	

Tabela 1.5: Note que a soma dos resíduos é nula devido a definição da média.

Em seguida, elevamos os resíduos ao quadrado.

$(-0,57)^2=0,32$	0,14	0,12	0,036	0,032
0,026	0,026	0,022	0,012	0,012
0,008	0,0036	0,0009	0,0001	0,0001
0,0004	0,0025	0,0064	0,01	0,0144
0,0169	0,0256	0,289	0,0324	0,04
0,0529	0,09	0,1156	0,16	

Tabela 1.6: A variância é o resíduo quadrático médio. Note que a variância, assim como a média é sensível a poucos dados muito afastados da média. Neste caso a variância será VAR=0,047. O desvio padrão será DP=$\sqrt{0,047}=0,22$.

No experimento de Cavendish podemos afirmar que a precisão de **cada** medida é de $0,22$, para mais ou para menos, e que a nossa melhor estimativa no momento para o valor da densidade da Terra é $\overline{x}=5,45$. Como já comentamos, a repetição de medidas irá aumentar a precisão do resultado final num processo de cancelamento de erro. Especificamente, o erro irá ser reduzido em $1/\sqrt{n-1}$ onde n representa o número de medidas. Em nosso exemplo o erro final será

$0{,}22/\sqrt{28} = 0,04$. A densidade da Terra que resulta do experimento pode ser finalmente divulgada como $\overline{x} = 5,45 \pm 0,04$.

1.3 Exercícios

1.3.1 Comparando Populações

Uma das principais aplicações da Estatística é a comparação de amostras provenientes de duas populações diferentes. O objetivo neste tipo de tarefa é a determinação do efeito de variáveis que estão sendo controladas sobre a variação de medidas. Mulheres que amamentam secretam cálcio no leite. Parte deste cálcio pode ser proveniente de seus próprios ossos, assim algumas mulheres podem experimentar perdas ósseas durante este período. Com o objetivo de medir as perdas ósseas, pesquisadores mediram o percentual de mudança do conteúdo ósseo em três meses em 20 mulheres em fase de amamentação e 10 mulheres de idades semelhantes que não estavam nem lactantes nem grávidas. Os dados são tabelados abaixo.

lactantes					não-lactantes				
-4,7	-2,5	-4,9	-2,7	-0,8	2,4	0,0	0,9	-0,2	1,0
-5,3	-8,3	-2.1	-6,8	-4,3	1,7	2,9	-0,6	-1,6	-2,2
2,2	-7,8	-3,1	-1,0	-6,5					
-1,8	-5,2	-5,7	-7,0	-0,3					

Tabela 1.7: Percentual de mudança do conteúdo ósseo em três meses de mulheres lactantes e de outras mulheres.

Mulheres em fase de amamentação apresentam mais perdas ósseas do que outras mulheres?

A forma mais direta de procedermos consiste em: 1. organizar os dados em um rol; 2. calcular o sumário de 5 números e construir um box plot.

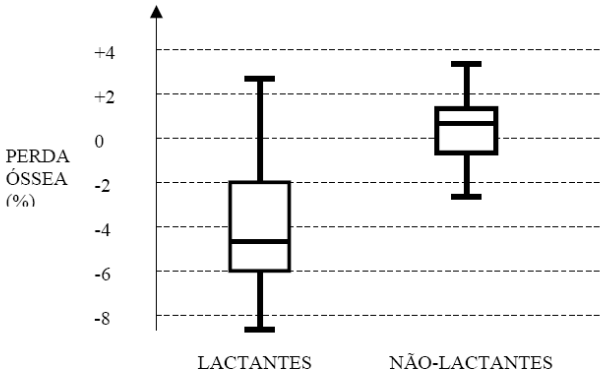
(a) Os dados organizados ficam

Mulheres em fase de amamentação					Outras mulheres				
-8,3	-7,8	-7,0	-6,8	-6,5	-2,2	-1,6	-0,6	-0,2	0,0
-5,7	-5,3	-5.2	-4,9	-4,7	0,9	1,0	1,7	2,4	2,9
-4,3	-3,1	-2,7	-2,5	-2,1					
-1,8	-1,0	-0,8	-0,3	2,2					

São dois os sumários possíveis: média e desvio padrão e os 5 números (min, max,mediana,q1,q3). Optamos pelos sumário de 5 números.

Mulheres em fase de amamentação					Outras mulheres				
MIN	Q1	MEDIANA	Q3	MAX	MIN	Q1	MEDIANA	Q3	MAX
-8,3	-6,1	-4,5	-1,95	2,2	-2,2	-0,6	0,45	1,7	2,9

(c) A representação mais adequada é o Box plot.

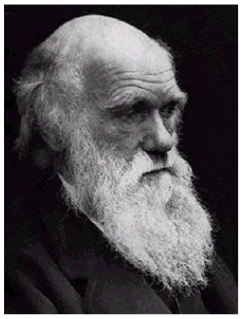


(c) O Box plot apresenta evidência de maior concentração de perdas em mulheres lactantes, com a mediana e terceiro quartis negativos. Uma conclusão mais definitiva dependeria de um teste de hipóteses para diferenças entre médias.

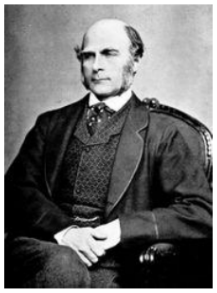
Figura 1.11: Solução para o exercício.

1.3.2 Darwin e Galton

Em seu livro 1876, *The Effect of Cross-and Self-fertilization in the Vegetable Kingdom*, Darwin descreveu uma série de experimentos projetados para demonstrar que a fertilização cruzada contribuiria para produzir plantas com crescimento mais vigoroso do que aquele observado em plantas auto-fertilizadas. Para a análise estatística dos dados Darwin consultou seu primo Francis Galton um dos pioneiros da bioestatística (e da eugenia).



Charles Darwin
(1809-1882)



Francis Galton
(1822-1911)

Os dados a baixo representam as alturas finais em polegadas plantas provenientes de pares de sementes de mesma idade. Em um tratamento (Cross) a fertilização foi cruzada, no outro (Self) houve auto-fertilização. Há evidência de diferença entre os tratamentos de fertilização?

Cross	*Self*
23,5	17,4
12,0	20,4
21,0	20,0
22,0	20,0
19,1	18,4
21,5	18,6
22,1	18,6
20,4	15,3
18,3	16,5
21,6	18,0
23,3	16,3
21,0	18,0
22,1	12,8
23,0	15,5
12,0	18,0

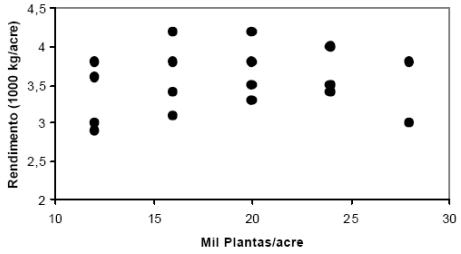
1.3.3 Eficiência Agrícola

A introdução de técnicas estatísticas é em grande medida responsável pela revolução verde na segunda metade do século 20. Um típico problema em agricultura é a determinação do rendimento de uma lavoura em função de características do solo ou de plano de plantação. O rendimento em kg de uma plantação de milho depende da taxa de plantas por acre. A tabela a seguir mostra o rendimento de para várias taxas de plantas por acre:

Mil Plantas/acre	Rendimento (1000 kg/acre)			
12	3,8	2,9	3,0	3,6
16	4,2	3,1	3,4	3,8
20	4,2	3,3	3,5	3,8
24	3,4	3,5	4,0	
28	3,0	3,8		

(1) Represente os dados da tabela graficamente. (2) Calcule o rendimento médio e o desvio padrão para cada taxa de plantas por acre. (3) Represente a média e o desvio padrão de forma gráfica (pontos com barras de erro). (5) Qual taxa de plantas por acre você sugeriria com base nas análises acima ?

(1)



(2)

Mil Plantas/Acre	*Rendimento médio*	*DP*
12	3,3	0,44
16	3,6	0,48
20	3,7	0,39
24	3,6	0,32
28	3,4	0,57

(3)

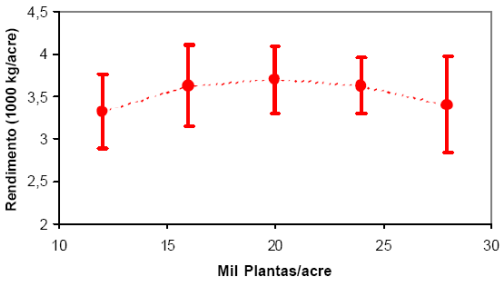


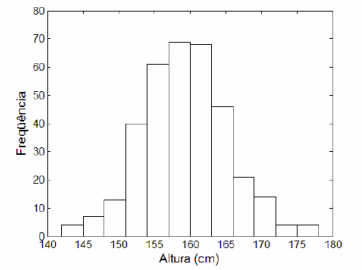
Figura 1.12: Solução para o exercício 3.

1.3.4 Média e Mediana

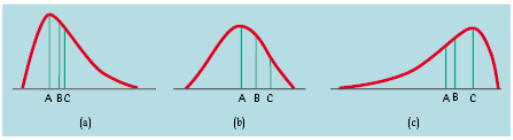
1. Em uma pesquisa envolvendo alunos formados em uma universidade particular a média e a mediana da renda mensal foram mensuradas. Os resultados (não necessariamente na ordem certa) foram R$ 3000,00 e R$

4000,00. Qual seria a média e qual seria a mediana? Justifique sua resposta descrevendo o raciocínio utilizado.

2. O histograma abaixo mostra a distribuição de alturas em centímetros de uma particular população de mulheres idosas. Marque sobre o gráfico a localização aproximada da média. Estime visualmente o valor e marque sobre o gráfico o desvio padrão desta população.



3. Nas figuras a seguir identifique qual das marcações é a média, qual a mediana e qual a moda.



1.3.5 Avaliação de Desempenho

Outra aplicação central de técnicas estatísticas é a avaliação de desempenho quando indicadores apresentam flutuações aleatórias. Os dados abaixo representam as cestas por partida de 2 jogadores de basquete, contadas por totais de pontos marcados. Apresente os resultados de forma de gráfica e compare os resultados dos dois jogadores.

Jogador 1	Jogador 2
15	50
45	12
32	46
16	10
30	20
90	0
26	52
52	44
34	10
40	45

1.3.6 Diagnóstico

Representações gráficas e Estatística podem ser empregadas para o detecção de problemas em áreas variadas, por exemplo, na gestão de negócios. O dono de uma venda desconfia que um de seus vendedores tem aconselhado seus fregueses a realizarem suas compras em outro estabelecimento. Para testar sua hipótese, ele contou o número de vendas realizadas a cada dia como função das horas que aquele funcionário passou atendendo os clientes. O estabelecimento ficou aberto durante o mesmo período de tempo em todos os dias, com outros funcionários responsáveis pelas vendas quando o funcionário suspeito não estava presente. Baseado apenas nos dados recolhidos abaixo para 9 dos dias em que as medidas foram feitas, verifique se as suspeitas do dono da venda são razoáveis.

Horas trabalhadas	Peças vendidas
4	20
5	50
2	60
8	20
5	40
7	60
1	80
2	60
7	10

1.3.7 Decisão

A Estatística fornece uma série de ferramentas para decisão utilizando informação quantitativa como no exemplo que segue. O dono de uma venda desconfia que um de seus vendedores tem aconselhado seus fregueses a realizarem suas compras em outro estabelecimento. Para testar sua hipótese, ele contou o número de vendas realizadas a cada dia como função das horas que aquele funcionário passou atendendo os clientes. O estabelecimento ficou aberto durante o mesmo período de tempo em todos os dias, com outros funcionários responsáveis pelas vendas quando o funcionário suspeito não estava presente. Baseado apenas nos dados recolhidos abaixo para 9 dos dias em que as medidas foram feitas, verifique se as suspeitas do dono da venda são razoáveis.

Transporte público	Automóvel
28	29
29	31
32	33
37	32
33	34
25	30
29	31
32	32
41	35
34	33

1.4 A Distribuição Normal

A descrição completa de fenômenos aleatórios é fornecida pela distribuição de probabilidade da variável aleatória que pode ser estimada em uma amostra pela construção de um histograma. Alguns tipos de histograma aparecem de forma recorrente e podem ser justificados por modelos matemáticos para processos aleatórios. Uma das formas mais genéricas é a distribuição em forma de sino conhecida como *distribuição normal* ou *distribuição gaussiana.* Galton formulou um experimento que nos mostra visualmente de que forma emerge a distribuição normal. A tábua de Galton (também conhecida como *quincunx*) está representada na figura a seguir.

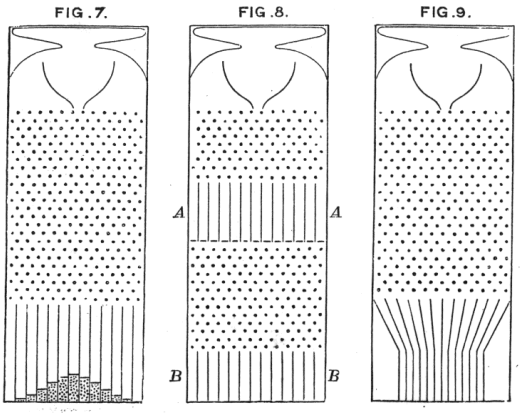


Figura 1.13: Quincunx de Galton em três variantes. As bolinhas caem até as calhas da extremidade inferior colidindo de forma aleatória com os pinos no caminho. Ao aumentarmos o tamanho da tábua observamos a formação do histograma em forma de sino.

Para entendermos um pouco dos elementos necessários para a formação da distribuição gaussiana podemos tentar construir um modelo matemático para o funcionamento do quincunx. Uma bolinha em queda se choca com um dos pinos da tábua e tem chance igual de ir para direita ou para esquerda a partir de sua posição inicial no centro. que identificamos por $S_0 = 0$. Se for para a direita diremos que ela se deslocou a distância positiva X_1em seu primeiro passo. Se de deslocar para a esquerda diremos que a distância é negativa. Sucessivamente a bolinha irá se chocar com pinos nas fileiras subseqüentes se deslocando, respectivamente, X_2 , X_3e assim por diante até X_n. A distância total a partir do centro no final irá definir a divisão da calha na qual a bolinha irá ser depositada e será a soma dos deslocamentos ao atravessar cada fileira:

$$S_n = X_1 + X_2 + X_3 + \cdots + X_n$$

A distribuição da variável aleatória $S_n$tende ao formato do sino a medida que n cresce. Este é o conteúdo do *Teorema do Limite Central.* Basta que tenhamos uma soma grande de números aleatórios independentes para darmos origem à uma distribuição normal. O interesse da distribuição é, no entanto, muito mais amplo, pois, podemos utilizar o mesmo modelo matemático do

quincunx para descrevermos fenômenos mais corriqueiros. Por exemplo, alturas de pessoas ou tamanho de folhas.

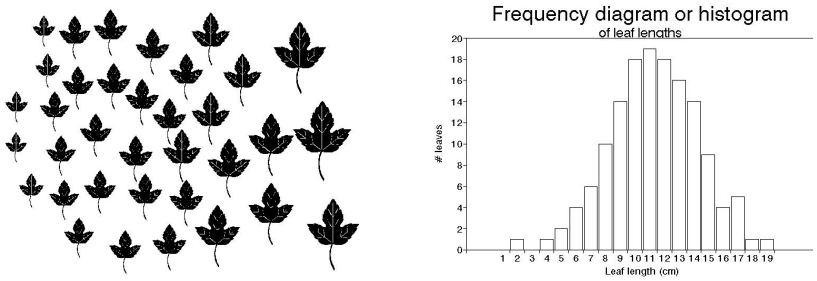


Figura 1.14: Histograma para o comprimento de folhas. Podemos imaginar que exista um comprimento básico (médio) de folhas e que um grande número de fatores se some para produzir variações deste comprimento para cima ou para baixo, condições suficientes para emergência de uma distribuição normal.

A distribuição normal apresenta propriedades matemáticas que nos possibilitam uma interpretação direta do que são medidas típicas e do significado do desvio padrão.

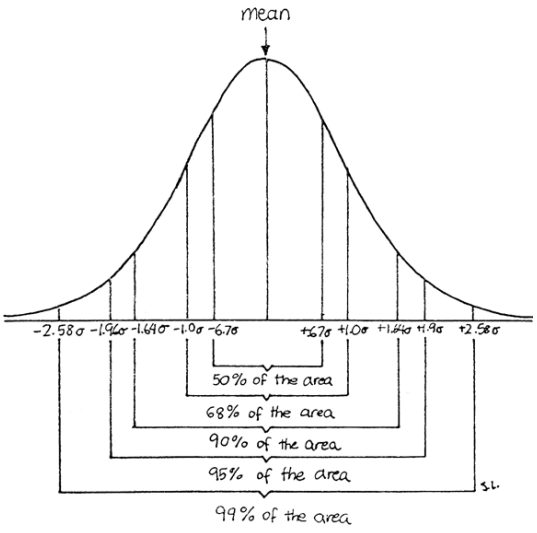


Figura 1.15: Lembrando que a área sob a densidade de probabilidade descreve o percentual de casos no intervalo. A distribuição normal é simétrica, a média (*mean)*, a mediana e a moda situam-se no centro. 50% dos dados ficam em um intervalo que vai da *média* - $0,67\sigma$ até a *média* + $0,67\sigma$. 68% dos dados estão no intervalo entre *média* - σ até a *média* + σ. 95% dos dados estão entre *média* - 2σ até a *média* + 2σ.

A distribuição normal acontece com muita freqüência, assim em geral é um bom começo supormos que o histograma de nossa amostra irá tender a uma distribuição normal conforme aumentamos seu tamanho. É claro que

essa suposição nem sempre estará correta pois há uma série de fenômenos cuja distribuição difere bastante de uma normal, no entanto em Estatística **sempre** trabalharemos com algum tipo de versão aproximada da realidade. Um exemplo de distribuição muito diferente de uma normal é a distribuição de rendas familiares.

Figura 1.16: Distribuição de renda familiar na Grande São Paulo e no Brasil. Note que a concentração de renda é maior no Brasil do que em São Paulo. As distribuições são assimétricas e a chance de grandes desvios da média é maior do que na distribuição normal.

1.4.1 Escore Z

A distribuição normal permite uma definição bastante específica de valores altos ou baixos. O que significa uma pessoa ser alta ou baixa, ou o que significa uma nota no vestibular alta ou baixa? Nossa espectativa inicial é que as notas do vestibular tenham distribuição normal. De fato isso ocorre nas disciplinas humanísticas, mas não em matemática ou física. Na figura a seguir mostramos um histograma das notas da primeira fase da FUVEST na área de exatas em 2007.

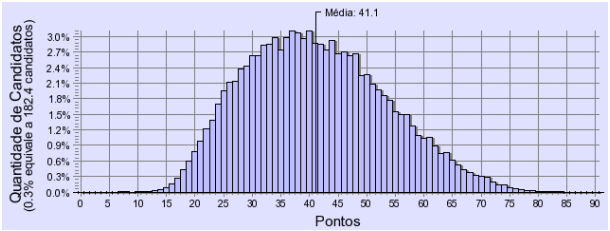


Figura 1.17: Apesar de claramente assimétricas (menos pessoas com notas altas do que baixas) em uma primeira aproximação as notas da FUVEST tem distribuição normal. A média do histograma é $\overline{x} = 41,1$ e o desvio padrão é $\sigma = 15$ (localize-os na figura). Como podemos julgar se fomos bem ou mal no vestibular?

Para sabermos se fomos bem ou mal no vestibular podemos usar as pro-

priedades matemáticas da distribuição normal para definir o *escore Z*.

$z = \dfrac{x - \overline{x}}{\sigma}$. O escore Z indica quantos desvios padrão a observação x dista da média. Supondo que o histograma da variável aleatória tem a forma de uma distribuição normal (que é uma aproximação, mas uma boa aproximação como ponto de partida), de posse da estatística Z, podemos consultar uma tabela para sabermos o percentual de observações que estaria abaixo do valor observado.

Digamos que nossa nota no vestibular foi $x = 60$. Será que fomos bem ou mal no vestibular. O escore Z será $z = \dfrac{60 - 41,1}{15} = 1,26$. Qual o percentual de candidatos com nota mais baixa do que a que obtive? A tabela Z (ou da normal padrão) mostra a probabilidade de observações menores do que z, denotamos esta probabilidade por $\Phi(z)$(letra phi maiúscula). Na tabela da próxima página temos em cada linha os primeiros dois dígitos do valor de Z e em cada coluna o último dígito.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Figura 1.18: Tabela Z (normal padrão). A nossa nota tem valor z=1,26. Buscamos a linha 1,2 e a coluna 0,06 para acharmos $\Phi(1,26) = 0,8962$. Ou seja, nossa nota foi maior do que a de aproximadamente 89,62% dos candidatos. Podemos afirmar que fomos bem na prova.

O escore Z também permite que representemos graficamente detalhes de

distribuições, por exemplo, veja o gráfico de referência para o peso de meninos pela idade.

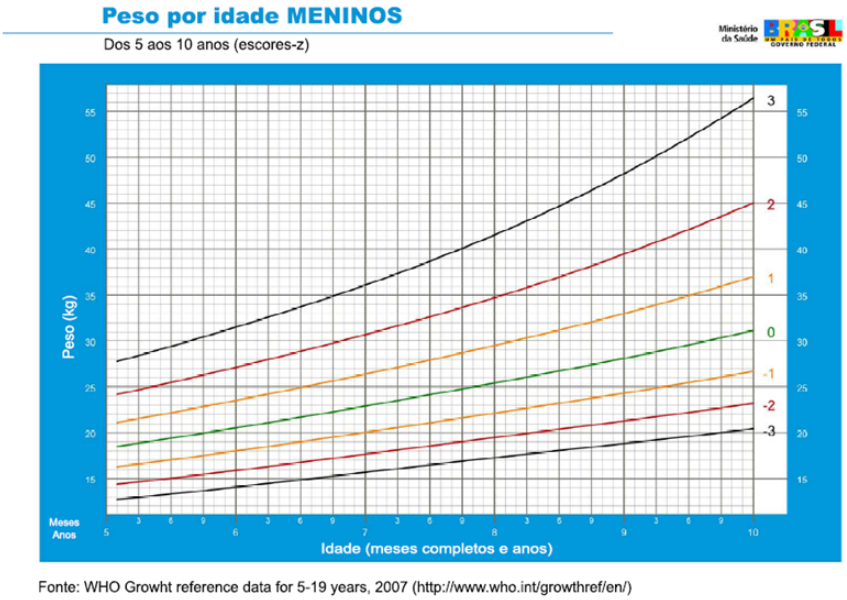


Figura 1.19: Cada linha da figura corresponde aos pesos com um certo escore Z. Assim podemos dizer que 97,7% (z=2) das crianças de 9 anos pesam menos que 39,5 kg (localize este ponto no gráfico). Os escores Z permitem que um gráfico apenas transmita muito mais informação.

1.5 Estimação Estatística

Frequentemente desejamos medir alguma quantidade em uma população inteira com recursos e tempo escassos. As técnicas de Estatística permitem grande economia de trabalho nessas circunstâncias. O procedimento denominado *estimação estatística* consiste da extração de informação relativa população através da observação de uma amostra dessa população. Como ilustração suponha que desejemos realizar uma pesquisa de opinião. Cada questão sendo respondida na forma de uma escala de 1 a 5. Queremos avaliar a nota média em uma população. A esta média populacional denominamos *parâmetro populacional*. Se tivéssemos recursos suficientes poderíamos colher respostas de toda a população realizando um *censo*, no entanto, utilizaremos as técnicas da estimação estatística para obtermos uma *estimativa* da escala média utilizando uma amostra (veja a figura).

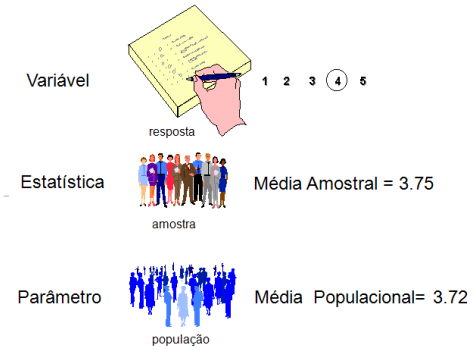


Figura 1.20: Nossa variável aleatória corresponde a escala de pesquisa. Queremos determinar o valor de um parâmetro populacional, no caso, a resposta média aos questionários. Como não temos recursos suficientes para realizarmos um censo que permitiria a determinação direta do parâmetro populacional, colhemos uma amostra. Qualquer cálculo que façamos utilizando dados de uma amostra é denominado uma *estatística*. Assim calculamos a resposta média na amostra como uma estatística que sirva como um *estimador* do parâmetro populacional.

Ao escolhermos uma amostra particular corremos o risco de obtermos uma estimativa muito diferente do parâmetro populacional. Não temos como saber se nossa amostra está próxima ou não do parâmetro populacional a menos que realizemos um censo. A teoria de probabilidades permite, no entanto, que determinemos um intervalo que contenha o valor do parâmetro populacional com probabilidade definida. Em outras palavras, conseguimos fazer uma afirmação sobre o valor do parâmetro populacional com risco definido de estarmos errados. Para visualizarmos como isso pode ser possível, comecemos por um experimento mental (ou *Gedankenexperiment*, como Einstein costumava denominar). Imaginemos o que aconteceria se coletássemos várias amostras e determinássemos estimativas utilizando cada uma delas. Os resultados certamente variariam, esta variação seria maior a medida que as amostras diminuíssem de tamanho e menor a medida que aumentassem. No limite, com amostras do tamanho de população teríamos vários censos e a variação seria nula.

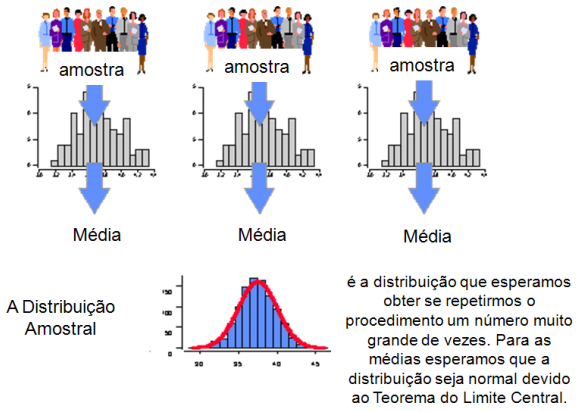


Figura 1.21: O Teorema do Limite Central garantirá que as estimativas de médias obtidas utilizando amostras suficientemente grandes tenham distribuição normal. A distribuição das médias é denominada *distribuição amostral.* A média da distribuição amostral é igual à média populacional μ(letra mü, pronunciada articulando-se um "u" e pronunciando um "i"). O desvio padrão é igual a $\frac{\sigma}{\sqrt{n}}$, onde σ é o desvio padrão da população e n é o tamanho da amostra.

Note que **não** conhecemos os parâmetros populacionais. Suponhamos, como ponto de partida, que conheçamos o desvio padrão populacional σ e o tamanho da amostra n. Obtemos uma estimativa para a média populacional $\hat{\mu}$. Queremos agora definir um intervalo que nos forneça 95% de chance de conter, caso fizéssemos um censo, o valor populacional μ. Como poderíamos definir tal intervalo ? Suponha que definamos o intervalo como $[\hat{\mu} - 2\frac{\sigma}{\sqrt{n}}, \hat{\mu} + 2\frac{\sigma}{\sqrt{n}}]$, a estimativa irá variar de acordo com a distribuição amostral. A chance da média populacional estar fora do intervalo é igual a chance de que o z escore de $\hat{\mu}$ seja maior que 2 ou menor que -2. Ou seja, a chance é de cerca de 5%. Assim podemos dizer que o intervalo definido tem chance de 95% de conter o valor desconhecido da média populacional.

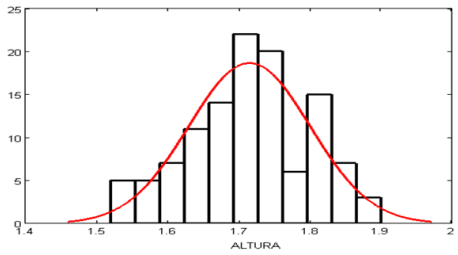


Figura 1.23: Histograma obtido ao medirmos alturas em uma amostra de tamanho $n = 115$. Nossa expectativa inicial é de um histograma em forma de curva normal, assim calculamos a média e o desvio padrão de nosso conjunto de dados e com isso desenhamos uma curva normal sobre o histograma para visualizarmos a semelhança. A média na amostra é $\overline{x} = 1,71m$ e o desvio padrão é de $\sigma = 0,09m$. Sabemos então que cerca de 68% dos dados estarão no intervalo $[1,62, 1,80]$ e cerca de 95% dos dados estarão entre $[1,53, 1,89]$.

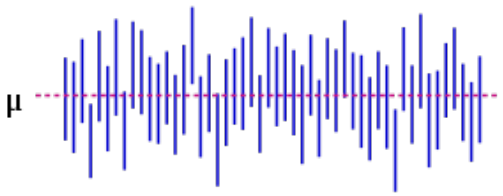


Figura 1.22: Cada intervalo de confiança na figura é resultado de uma amostragem. Na prática coletaremos apenas uma amostra, em 95% dos casos μ estará dentro de nosso intervalo. O risco de estarmos errados estará sempre presente (sinto informá-los !).

Para darmos um exemplo concreto do processo de estimação, digamos que desejemos determinar a altura média em uma população. Para isso coletamos uma amostra de tamanho $n = 115$ tomando o cuidado para que esta amostra não seja viciada, ou seja, para que o critério que usarmos na escolha da amostra não tenha relação com a variável que queremos medir (assim, por exemplo, não iremos medir 115 jogadores de basquete ou 115 jóqueis !). O histograma obtido aparece na figura a seguir:

Um *estimador sem viés* é uma estatística que tende ao parâmetro a medida que aumenta o tamanho da amostra. Quanto mais rápido o valor do estimador tender ao valor correto do parâmetro populacional mais *eficiente* ele será. Um estimador sem viés e eficiente para a média populacional será a média amostral

usual:

$\hat{\mu} = \overline{x} = \dfrac{x_1 + x_2 + . \cdots + x_n}{n}$ o estimador para a média populacional coincide com a média amostral usual.

Em seguida precisamos avaliar a largura do intervalo de confiança. Para isso precisamos de uma estimativa para o desvio padrão populacional $\hat{\sigma}$. Um estimador eficiente e sem viés para o desvio padrão amostral será a raiz quadrada da *variância amostral,* que é a variância usual dividida por $n-1$ ao invés de n.

$$\text{VAR}_{amostral} = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

O estimador para o desvio padrão populacional será então:

$\hat{\sigma}$=$\sqrt{VAR_{amostral}}$. Note que esta estimativa está relacionada com o desvio padrão calculado na amostra da seguinte maneira:
$\hat{\sigma}$=DP$\sqrt{\dfrac{n}{n-1}}$, ou seja, o desvio padrão usual tenderá a subestimar o desvio padrão populacional. Nas planilhas eletrônicas há sempre as duas versões. O desvio padrão populacional é nosso DP. O desvio padrão amostral é nosso $\hat{\sigma}$.

Em nosso exemplo calculamos $\hat{\mu} = \overline{x} = 1,71$ e DP=0,09. Para obtermos a estimativa do desvio padrão populacional efetuamos a correção acima de forma que $\hat{\sigma}$=DP$\sqrt{\dfrac{115}{114}} = 0,09 \times 1,004 = 0,09$ (note que para amostras grandes não

faz diferença se usamos VAR ou $VAR_{amostral}$. O intervalo de confiança será :

$\hat{\mu} \pm$ (Fator de Confiança) $\times \frac{\hat{\sigma}}{\sqrt{n}}$. Aqui estamos utilizando uma estimativa para o desvio padrão populacional $\hat{\sigma}$ no lugar do valor correto σ. É claro que esta estimativa terá ela própria um erro, que estamos ignorando. Para amostras não muito pequenas a aproximação é boa. Para amostras pequenas há uma versão avançada deste cálculo que leva em conta estes erros (utilizando as chamadas distribuições T de Student). O Fator de Confiança irá depender do risco de estar errados que aceitamos correr. Quanto menos risco aceitarmos menos informativa será nossa estimativa. Se quisermos 50% de chance de estarmos errados, escolheremos o fator de 0,67, se quisermos 32% de chance de erro, escolheremos o fator como 1. Para 5% de chance de erro, o fator será 2 e para 0,3% será 3.

Continuando nosso cálculo teremos, para 95 % de confiança

$$\hat{\mu} = 1,71 \pm 2 \times \frac{0,09}{\sqrt{115}} = 1,710 \pm 0,017$$

Aqui a convenção é usar dois algarismos significativos quando o primeiro significativo for 1 ou 2, e um algarismo significativo quando for maior que 2. Assim obtivemos 0,017 e acrescentamos um algarismo significativo escrevendo 1,710. Outra foram alternativa de escrevermos o resultado é $\hat{\mu} = 1,710(17)$.

1.5.1 Determinação do tamanho de uma amostra

Podemos usar nossas técnicas para avaliar intervalos de confiança de forma inversa para obtermos o tamanho da amostra n em termos do erro desejado. Quanto menor o erro desejado, maior será o investimento em coleta de dados que precisaremos fazer. O tamanho da amostra será:

$n = \left[\frac{\text{(Fator de Confiança)} \times \hat{\sigma}}{Erro}\right]^2$. Aqui um nível de erro estatístico aceitável tem que ser fixado por nós. a estimativa para o desvio padrão populacional tem que ser obtida em um estudo piloto, extraída de estudos anteriores semelhantes ou simplesmente fixada a priori de forma conservadora. O fator de confiança também é fixado por nós, sendo tradição (não uma regra justificável) fixar-se 2 para termos 95% de confiança.

Digamos então que queremos saber o tamanho da amostra necessária para estimarmos alturas com erro de 5 cm para cima ou para baixo. Fixemos $\hat{\sigma} = 0,10$. Teremos então

$$n = \left[\frac{2 \times 0,1}{0,05}\right]^2 = 16$$

Assim, uma amostra com 16 pessoas seria suficiente apra atingirmos a precisão desejada. Até agora apenas comentamos o caso de variáveis quantitati-

vas, frequentemente, no entanto, desejamos estimar percentuais de ocupação de classes em variáveis qualitativas (por exemplo, no caso de eleições ou pesquisas de opinião). Neste caso teremos a seguinte expressão:

Primeiro calculamos $n^\star = \left[\frac{1}{Erro}\right]^2$. A seguir obtemos o tamanho da amostra como $n = \frac{n^\star}{1 + \frac{n^\star}{N}}$, onde N é o tamanho da população.Esta estimativa para o tamanho de amostra é pessimista, ou seja, é possível que o erro que encontrarmos após a pesquisa seja menor do que o erro que especificamos inicialmente.

Digamos então que queremos saber se uma determinada população com $N = 5000$ pessoas é contra ou à favor de certa idéia. Gostaríamos de avaliar os percentuais contra ou à favor com erro de 5% para cima ou para baixo. Assim temos $n^\star = (1/0,05)^2 = 400$ e $n = 400/(1 + 0,08) = 370$. Ou seja, precisaríamos entrevistar 370 pessoas em uma estimativa pessimista.

1.6 Exercícios

1.6.1 Medidas Experimentais

Experimentos jamais fornecem apenas um valor quando repetimos medidas muitas vezes. Quanto mais repetirmos as medidas maior tenderá a ser a precisão de nosso resultado. Um experimento para determinação da taxa metabólica masculina. O consumo de calorias em um dia foi medido em uma amostra de $n = 7$ indivíduos. As medidas obtidas em cal/dia estão listadas na tabela abaixo:

1792	1666	1392	1642	1460	1867	1439

Estime o consumo médio na população utilizando um intervalo de confiança de 95%. (*Siga os passos usuais: 1. calcule a média aritmética para estimar a média populacional; 2. calcule a variância amostral e o desvio padrão amostral para estimar o desvio padrão populacional; 3. calcule o intervalo de confiança multiplicando o fator de confiança adequado (=2) pela estimativa do desvio padrão populacional e divida pela raiz do tamanho da amostra menos um.* Resposta: $\hat{\mu} = 1608 \pm 150$).

1.6.2 Teste de Hipóteses

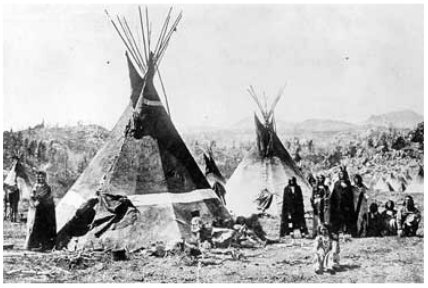


Figura 1.24: Os Shoshoni , tribos de nativos americanos do Wyoming, costumavam decorar seus objetos de couro com retângulos. Em *Lowie's Selected Papers in Anthropology (1970) Dubois,C. (ed)* levantou-se a hipótese de que os Shoshoni estariam empregando a razão áurea em suas decorações.

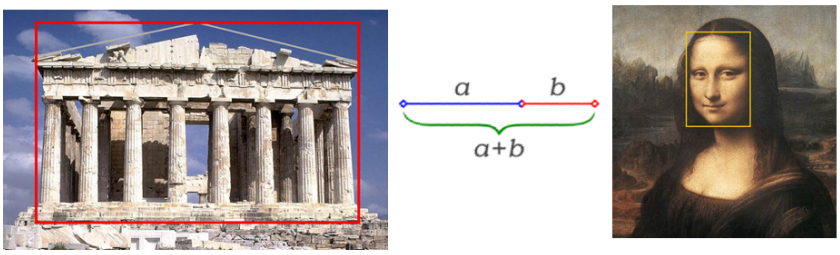


Figura 1.25: A razão áurea é definida como $\varphi = \frac{a}{b} = \frac{a+b}{a}$=1,6180..... A razão áurea é utilizada para dimensionar a relação entre largura e altura em arquitetura em artes plásticas correspondendo à razão de aspecto $\frac{1}{\varphi} = 0,6180....$

Gostaríamos de testar a hipótese do uso da razão áurea pelos Shoshoni, para isso foram colhidos os seguintes dados para a razão de aspecto de molduras..

0.693	0.662	0.690	0.606	0.570
0.749	0.672	0.628	0.609	0.844
0.654	0.615	0.668	0.601	0.576
0.670	0.606	0.611	0.553	0.933

1. Construa um diagrama de ramo e folhas para visualizar a distribuição dos dados. 2. Construa um histograma. 3. Calcule o sumário de 5 números. 4. Estime a razão de aspecto média na população com um intervalo de confiança de 95%. A razão áurea seria compatível com as observações com 95% de confiança? (*A razão áurea está dentro do intervalo estimado?)*

1.6.3 Empate Técnico

Uma pesquisa pré-eleitoral foi conduzida utilizando-se uma amostra de tamanho $n = 5000$ de uma população de $N = 10$milhões de eleitores. O candidato A recebeu 44% das intenções de voto e o candidato B recebeu 41% da preferência. 12% do eleitorado declarou intenção de anular seu voto. Podemos afirmar que o candidato A ganharia as eleições?

1.6.4 Avaliando uma Medida em Relação à uma População

O nível sanguíneo de colesterol é importante devido aos riscos de doença coronária associados. Para meninos de 14 anos média populacional é estimada em $\hat{\mu} = 170$ mg/dl. com desvio padrão de $\hat{\sigma}$=30 mg/dl. Casos acima de 240 mg/dl requerem cuidados médicos especiais. Qual seria o percentual esperado de meninos de 14 anos com necessidade de tratamento?