Sumário

1 Análise exploratória

Para poder responder as questões desta primeira parte você deve leer o Capítulo 1 do livro [4], disponível na biblioteca. O objetivos desta primeira parte são

- introduzir algumas técnicas elementares para explorar dados tais como tabelas de frequência, gráficos de barras, histogramas, gráficos de caixa (*boxplots*) e diagramas circulares.

- introduzir a análise exploratória utilizando a linguagem R. As técnicas mencionadas no item anterior são implementados em R utilizando respectivamente as funções `table`, `barplot`, `hist`, `boxplot` e `pie`.

A maneira de exemplo, considere os dados da Tabela 1.1 em [4], os quais podem ser importados em R com função `read.table`,

```
dt <- read.table("http://dcm.ffclrp.usp.br/~rrosales/aulas/escola.txt",
        header=TRUE)
attach(dt)
```

A primeira instrução guarda os dados na variável `dt`, em uma estrutura de dados conhecida como *data frame*. A argumento `header=TRUE` permite identificar cada coluna pelo seu nome (veja a descrição destas ultimas nas paginas 5-6 em [4]), e a instrução `attach(dt)` permite acesar os dados de cada coluna diretamente pelo nome da coluna. Os seguintes comandos reproduzem respectivamente as Figuras 1.4, 1.5, 1.6, e 1.8 em [4],

```
pie(table(Toler))
barplot(table(Idade), xlab="Idade", ylab="Frequencia")
hist(Peso, xlab="Peso", ylab="Densidade", prob=TRUE)
boxplot(Peso, ylab="Peso")
```

A Figura 1.9 é reproduzida pelos comandos

```
boxplot(dt[dt$Sexo == "F", 6], dt[dt$Sexo == "M", 6], names=c("F", "M"))
```

Sugiro você digitar `help(*)`, substituindo `*` por qualquer uma das funções utilizadas acima para aprender a utilizar melhor cada uma (`q` volta a linha de comando).

Exercício 1. Quince pacientes de uma clínica de ortopedia foram entrevistados quanto ao número de meses previstos de fisioterapia, se haverá (S) ou não (N) sequelas após o tratamento e o grau de complexidade da cirugia realizada: alto (A), médio (M) ou baixo (B). Os dados são apresentados na seguinte Tabela:

Paciente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Fisioterapia	7	8	5	6	4	5	7	7	6	8	6	5	5	4	5
Sequelas	S	S	N	N	N	S	S	N	N	S	S	N	S	N	N
Cirugia	A	M	A	M	M	B	A	M	B	M	B	B	M	M	A

(i) Para o grupo de pacientes que não ficaram com sequelas, faça um gráfico de barras para a variável Fisioterapia. Compare com o grafico do grupo de pacientes com sequela. Você acha que a variável Fisioterapia se comporta de modo diferente? (ii) Considere os grupos determinados pela variável Cirugia e estude o comportamento de Fisioterapia destes grupos. Existem diferenças? A tabela acima pode ser carregada digitando

```
dt <- read.table("http://dcm.ffclrp.usp.br/~rrosales/aulas/orto.txt",
        header=TRUE)
```

Exercício 2. Faça o Exercício 4 da Seção 1.2 em [4].

Exercício 3. Faça o Exercício 5 da Seção 1.2 em [4].

Exercício 4. Inicie uma sessão do R e carregue os dados `cancer.txt` utilizando o comando

```
dt <- read.table("http://dcm.ffclrp.usp.br/~rrosales/aulas/cancer.txt",
        header=TRUE)
```

O arquivo `cancer.txt` contem os dados de uma pesquisa relativos a incidência de câncer, organizados em nove colunas. Os nomes de cada uma podem ser determinados ao digitar

```
names(dt)
```

`Ident`: identifica o paciente; `Grupo`: determina o diagnóstico sendo 1=falso negativo: pacientes diagnosticados como não tendo a doença quando na verdade tinham, 2=negativo:

diagnosticados como não tendo a doença quando de fato não tinham, 3=positivo: diagnosticado corretamente como tendo a doença, 4=falso positivo: diagnosticados como tendo a doença quando na verdade não tinham; `Idade`: idade do paciente; `AKP`: espectro químico da análise do sangue-alkaline phospatase; `P`: concentração do fosfato no sangue; `LDH`: enzima lactate dehydrogenase; `ALB`: albumina; `N`: nitrogênio na uréia; `GL`: glicose. (i) Uma afirmação feita por alguns médicos é a de que o grupo dos falso-positivos é mais jovem do que o dos falso-negativos. Para os dados dessa pesquisa, o que você diria a respeito? Justifique a sua resposta baseandose em gráficos e tabelas de frequência.

Uma maneira de escolher as filas contendo os dados do grupo falso positivo é

```
G4 <- dt[dt$Grupo == '4',]
```

Podemos agora escolher os valores da `Idade` deste subconjuto determinados pela terceira coluna

```
Idade.G4 <- G4[,3]
```

Claro, poderiamos também ter feito isto em um passo,

```
Idade.G4 <- dt[dt$Grupo == '4',3]
```

Outra forma flexivel para escolher subconjuntos dos dados esta baseada na função `subset`. Por exemplo,

```
Idade.G4 <- subset(dt, Grupo > 3, Idade)
```

seleciona a coluna `Idade` dos pacientes do grupo 4.

Exercício 5. Na linha de produção de uma grande montadora de veículos, existem 7 verificações do controle de qualidade. Sorteamos alguns dias do mês e anotamos o número de OKs recibidos pelos veículos produzidos nesses dias, i.e., em quantos dos controles mencionados o automóvil foi aprovado. Os resultados foram $((x, y)$, x =número de aprovações, y =freqüência): (4, 126), (5, 359), (6, 1685), (7, 4764). (i) Determine a média, moda e mediana do número de aprovações por automóvel produzido. (ii) Calcule a variância da amostra. (ii) Crie uma nova variável "reprovações", indicando o número de verificações não OKs no vehículo. Determine média, moda, mediana e variância dessa variável. Em geral, se uma amostra qualquer esta constituída pelas observações $z = (z_1, z_2, \ldots, z_n)$, então

$$\bar{z} = \sum_{i=1}^{n} z_i/n \qquad \text{média amostral}$$

seja $\tilde{z}_1 \leq \tilde{z}_2 \leq \cdots \leq \tilde{z}_n$ a amostra ordenada em forma crescente, então

$$m_d = \begin{cases} \tilde{z}_{(n+1)/2} & \text{se } n \text{ impar,} \\ \frac{1}{2}(\tilde{z}_{n/2} + \tilde{z}_{n/(2+1)}) & \text{se } n \text{ par} \end{cases} \qquad \text{mediana amostral}$$

$$m_o = \text{ valor mais frequente} \qquad \text{moda amostral}$$

$$\text{var}(z) = \sum_{i=1}^{n} (z_i - \bar{z})^2/n \qquad \text{variância amostral}$$

(iv) Cada reprovação implica em custos adicionais para a montadora, tendo em vista a necessidade de corrigir o defito apontado. Admitindo um valor básico de R$ 200,00 por cada item reprovado num vehículo, calcule a média e a variância da espesa adicional por automóvel produzido.

Exercício 6. Um hospital maternidade está planejando a ampliação dos leitos para recém nascidos. Um levantamento do número de dias que os bebes permanecerem no hospital forneceu os seguintes dados: 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 7, 7, e 8. (i) Determine a tabela de frequência e calcule a média, moda e mediana destes dados. (ii) Determine o desvio padrão. (iii) Dentre as medidas de posição calculadas em (i), quais delas seriam mais apropriadas para ressumir o conjunto de dados?

Exercício 7. Com os dados do **Exercício 4**: (i) obtenha as medidas de posição e variabilidade para as variáveis Idade e Glicose (GL), (ii) repeta o item (i) para cada tipo de diagnóstico. Compare as respostas obtidas.

2 Estimação pontual

2.1 Estimadores

Exercício 8. Foram sorteadas 15 famílias com filhos num certo bairro e observado o número de crianças de cada família, matriculadas na escola. Os dados foram 1, 1, 2, 0, 2, 0, 2, 3, 4, 1, 1, 2, 0, 0, e 2. Obtenha as estimativas correspondentes aos seguintes estimadores da média de crianças na escola nesse bairro,

$$\widehat{\mu}_2 = \frac{(X_1 + X_2)}{2}, \qquad \widehat{\mu}_3 = \bar{X}.$$

Qual deles é o melhor estimador da média e por quê?

Exercício 9. Seja X_1, X_2, X_3 uma amostra aleatória de uma população exponencial com média θ, isto é, $\mathbb{E}[X_i] = \theta, i = 1, 2, 3$. Cosidere os estimadores

$$\widehat{\theta}_1 = \bar{X}, \qquad \widehat{\theta}_2 = X_1, \qquad \widehat{\theta}_3 = \frac{X_1 + X_2}{2}.$$

(i) Mostrar que nenhum dos três estimadores é viesado. (ii) Qual dos estimadores tem menor variância? Lembrar que para o modelo exponencial $\text{Var}(X_i) = \theta^2$.

Exercício 10. (Este exercício tem implicações fundamentais para a estatística) Sejam X_1, X_2, ..., X_n variáveis aleatórias independentes e identicamente distribuídas com média μ e variância σ^2. Sejam

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{e} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

(i) Determine $\mathbb{E}[\bar{X}_n]$ e $\text{Var}(\bar{X}_n)$. (ii) Mostre que $\bar{X}_n$ é consistente para μ. (iii) Mostre que

$$S_n^2 = \frac{n}{n-1}\left\{\sum_{i=1}^{n} X_i^2 - n(\bar{X}_n)^2\right\}.$$

(iv) Calcule $\mathbb{E}[S_n^2]$. (v) Utilice os resultados em (iii) e em (iv) para mostrar que S_n^2 é consistente para σ^2. [Sugestão: utilice duas vezes a Lei dos Grandes Numeros.]

Exercício 11. Seja X_1, X_2, ..., X_n uma amostra de uma população com distribuição

$$f_X(x) = \frac{2x}{\theta^2}, \qquad 0 < x < \theta, \quad \theta > 0.$$

Verifique se $\widehat{\theta}_1 = \bar{X}$ e $\widehat{\theta}_2 = \max\{X_1, X_2, \ldots, X_n\}$ são não viciados para θ. (ii) Calcule e compare os EQM dos estimadores em (i). (iii) Faça um gráfico dos EQM em função de θ. Sugestão: utilice R para fazer o gráfico em (iii)[1]. Sugestão: suponha que $M = \max\{X_1, \ldots, X_n\}$ é a v.a. correspondente ao máximo das v.as $X_1, \ldots, X_n$. Para determinar $\mathbb{E}[M]$ é necessário determinar a densidade de M, f_M. Sob independência é simples verificar que $F_M(x) = \left[F_{X_1}(x)\right]^n$, logo $f_M(x) = dF_M(x)/dx$.

[1] O seguinte exemplo ilustra os passos necessários para graficar duas funções f e g no dominio $x \in [-2, 10]$:

```
x <- seq(-2,10,by=0.01)
f <- exp(-x)+1/abs(x-1)
plot(x,f, type="l", col="navy", ylim=c(-1,30), lwd=2)
g <- 3*sin(x^3)/(3-x) + 10
lines(x, g, col="sandybrown", lwd=2)
```

Exercício 12. Suponha que Y tem distribuição Binomial-(n, p). (i) Demostre que $\widehat{p} = y/n$ é um estimador não viesado para p. Calcule a variância de $\widehat{p}$.

Exercício 13. Seja $U_1, U_2, \ldots, U_n$ uma amostra de uma população com densidade uniforme no intervalo $[\theta, 1]$, $\theta > 0$, e seja $M = \min\{U_1, U_2, \ldots, U_n\}$. O estimador M é viciado para θ, embora este premite definir um estimador não viciado. Determine o estimador não viciado para θ baseado em M. Diga se este ultimo é (fracamente) consistente. Observe que se U é uniforme em $[\theta, 1]$, então U possui densidade definida pela função

$$f_U(x) = \begin{cases} 1/(1-\theta), & \text{se } x \in [\theta, 1], \\ 0, & \text{caso contrário.} \end{cases}$$

Exercício 14. Seja X uma variável aleatória Binomial com parâmetros n e p. Suponha que n seja conhecido porém p é desconhecido. Considere os estimadores para p,

$$\hat{p}_1 = \frac{X}{n}, \qquad \hat{p}_2 = \frac{X+1}{n+2}.$$

(i) Mostre que dado n, $\mathrm{EQM}(\hat{p}_2) < \mathrm{EQM}(\hat{p}_1)$ sempre e quando p assoma valores no intervalo

$$\left(\frac{1}{2} - \frac{\sqrt{(n+1)(2n+1)}}{2(2n+1)}, \frac{1}{2} + \frac{\sqrt{(n+1)(2n+1)}}{2(2n+1)}\right).$$

(ii) Determine o valor deste intervalo para $n = 1, 2, 3, 4$ e mostre que para n suficientemente grande, este é próximo de $(0.146, 0.854)$.

Exercício 15. Suponha que uma moeda possue probabilidade de resultar em cara igual a p, $0 < p < 1$. Com o objetivo de estimar θ, a probabilidade de obter duas caras em lançamentos sucessivos, a moeda é lançada $n \geq 2$ vezes. Se como resultado são observadas X caras, mostre que

$$\hat{\theta} = \frac{X(X-1)}{n(n-1)}$$

é um estimador não viciado para θ. Observe que X possui distribuição Binomial(n, p).

Exercício 16. Suponha que o tempo X necessário para que um novo CPU realice uma determinada tarefa seja exponencialmente distribuído com parâmetro θ, ou seja, a sua densidade é

$$f(x; \theta) = \frac{1}{\theta} \exp\Big(-\frac{x}{\theta}\Big), \quad x > 0, \quad \theta > 0.$$

Considere uma amostra X_1, X_2, ..., X_n de X. (i) Mostre que a densidade do mínimo dos tempos da amostra, W, possui densidade

$$f(w; \theta) = \frac{n}{\theta} \exp\Big(-\frac{nw}{\theta}\Big), \quad w > 0, \quad \theta > 0.$$

(ii) Mostre que os estimador W é viciado para θ. (iii) Defina um estimador não viciado para θ baseado em W.

2.2 Projeto 1: função de distribuição empírica, histogramas

Este projeto tem varios objetivos: apresentar a noção de função de distribuição empírica de uma amostra e introducir os histogramas. Ambos estimadores são utilizados para inferir respectivamente a função de distribuição e a distribuição da população.

Suponhamos que $X_1, X_2, \ldots, X_n$ sejam variáveis aleatórias independentes e identicamente distribuídas, com função de distribuição F, e densidade f. A função de distribuição empírica da amostra $X_1, X_2, \ldots, X_n$ é definida como

$$\begin{aligned}\widehat{F}_{X_1,\ldots,X_n}(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}} = \frac{1}{n}\#\big\{i \in \{1,2,\ldots,n\} : X_i \leq x\big\} \\ &= \frac{1}{n}\big(\text{número de elementos na amostra} \leq x\big).\end{aligned}$$

(i) Veja o Apêndice a respeito da notação para os diferentes tipos de convergência de variáveis aleatórias a ser considerada no curso. Explique por que

$$\widehat{F}_{X_1,\ldots,X_n}(x) \xrightarrow{\mathbb{P}} F(x). \tag{1}$$

Este resultado justifica o emprego da distribuição empírica de uma amostra como um estimador da função de distribuição (cumulativa) da distribuição.

Considere agora $a = a_1 < a_2 < \ldots < a_m = b$, uma sequência de números reais (equidistantes), e os intervalos $A_k = (a_{k-1}, a_k]$ para $k = 2, \ldots, m$. Logo para $x \in A_k$ definimos

$$\begin{aligned}\widehat{h}_{X_1,\ldots,X_n}(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{a_{k-1} < X_i \leq a_k\}} \\ &= \frac{1}{n}\#\big\{\text{número de elementos na amostra} \in (a_{k-1}, a_k]\big\}.\end{aligned}$$

A função $\widehat{h}$ é conhecida como histograma. (ii) Mostre que para qualquer $x \in A_k$,

$$\widehat{h}_{X_1,\ldots,X_n}(x) \xrightarrow{\mathbb{P}} \int_{a_{k-1}}^{a_k} f(u)\, du. \tag{2}$$

[Sugestão: utilice (1)] Isto ultimo justifica a utilização dos histogramas como estimadores para as densidades. Inicie o R e digite

```
op <- par(mfrow = c(3, 1))
for (n in c(75,250,10000)) {h <- rnorm(n); hist(h,breaks=50, main=n);}
par(op)
```

A função `rnorm` gera uma amostra de tamanho `n` da densidade normal com média 0 e variância 1. A função `hist` calcula o histograma da amostra e o grafica. `breaks` determina o numero de intervalos nos quais sera avaliado o histograma e determina os extremos de cada um. (iii) Substitua `rnorm(n)` por: (a) `rgamma(n, 3, 5)` e (b) `rexp(n, 3)`, utilizando vários valores para n e variando o valor de `breaks` se for necessário. Em (a) você esta gerando uma amostra da distribuição gamma com parametros $\alpha = 3$, $\beta = 5$ e em (b) da exponencial com $\lambda = 3$. Comente os seus resultados.

2.3 Máxima verossimilhança

Exercício 17. Seja $X = X_1, X_2, \ldots, X_n$ uma amostra aleatória da uma população com densidade Gamma-(α, β), com $\alpha = 2$, e β desconhecido, isto é,

$$f(x) = \begin{cases} \dfrac{x\, e^{-x/\beta}}{\beta^2} & \text{se } x > 0, \\ 0 & \text{se } x \leqslant 0. \end{cases}$$

(i) Obtenha o estimador de máxima verosimilhança para β. (ii) Calcular $\mathbb{E}[\widehat{\beta}]$. É $\widehat{\beta}$ viciado para β?

Exercício 18. Determinada população é vitima de um surto de dengue. (i) Uma amostra aleatória de n pessoas é examinada. Baseado na amostra, determine o estimador de máxima verossimilhança para a proporção ν de pessoas infestadas pelo virus na população. (ii) Suponha que os indivíduos da população são examinados um a um até aparecer a primeira pessoa infestada. Seja T o número de pessoas examinadas. Se este procedimento é repetido n vezes, sejam T_1, T_2, ..., T_n o número de tentativas de cada vez. Qual é o estimador de máxima verossimilhança para ν baseado nesta amostra? É possível determinarmos qual dos procedimentos (i) ou (ii) é melhor para estimarmos ν?

Exercício 19. . Suponha que sejam realizados ensaios Bernoulli independentes, cada um com a mesma probabilidade de sucesso $0 < p < 1$ até que r sucessos sejam acumulados. Seja X o número total de ensaios requeridos, logo

$$P(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \qquad n = r, r+1, \ldots$$

Esta distribuição é conhecida como a distribuição Binomial Negativa com parâmetros r e p, e é utilizada para modelar a probabilidade de observarmos r expressos de um determinado gene em uma livraria de tamanho n. (i) Determine o estimador de máxima verossimilhança para p supondo que r seja conhecido. (ii) Suponha que em três livrarias com $r = 10$ foram observados os seguintes valores para n: $n_1 = 19$, $n_2 = 17$ e $n_3 = 17$. Calcule o valor do estimador em base a esta amostra. (iii) Suponha que outras três livrarias fornecem os dados $n_1 = 119$, $n_2 = 107$ e $n_3 = 121$. Determine o valor do estimador de máxima verossimilhança para p baseado nesta amostra. (iv) Compare os resultados em (ii) e (iii).

Exercício 20. Seja X_1, X_2, ..., X_n, uma amostra de uma população com distribuição $f_X(x) = \theta^x (1-\theta)^{1-x} \mathbf{1}_{\{0,1\}}(x)$, onde $0 \leq \theta \leq \frac{1}{2}$. (i) Encontre o estimador $\widehat{\theta}$ de máxima verossimilhança para θ. (ii) Calcule o EQM($\widehat{\theta}$), o erro quadrático médio de $\widehat{\theta}$. (iii) Diga se $\widehat{\theta}$ é (fracamente) consistente.

Exercício 21. Suponha que certa população X seja caracterizada pela distribuição: $P(X = 0) = \frac{2}{3}\theta$, $P(X = 1) = \frac{1}{3}\theta$, $P(X = 2) = \frac{2}{3}(1-\theta)$ e $P(X = 3) = \frac{1}{3}(1-\theta)$. (i) Determine o valor do estimador de verossimilhança para θ, $\hat{\theta}_{MV}$, se é considerada a seguinte amostra de X: 3, 0, 2, 1, 3, 2, 1, 0, 2, 1. O seguinte código em R grafica a função de verossimilhança para esta amostra.

```
logL <- function(theta) {
   2*(log(2/3)+log(theta))+3*(log(1/3)+log(theta))
   +3*(log(2/3)+log(1-theta))+2*(log(1/3)+log(1-theta))
}
theta <- seq(0,1,0.01)
plot(theta,logL(theta),ylab="verossimilhanca", xlab="theta",ty="l",lwd=2)
```

(ii) Diga se a estimativa para $\hat{\theta}_{MV}$ obtida em (i) atinge o máximo da função de verossimilhança.

Exercício 22. Seja $P(k)$ a distribuição do grau em uma rede complexa; ou seja, $P(k) = \mathbb{P}(X = k)$, $k \geq \mathbb{N}$, é a distribuição de probabilidade da variável aleatória X correspondente ao número de elos incidentes a um vétice (qualquer) da rede. Existem vários modelos probabilisticos os quais definem a 'topologia da rede' por exemplo

(1) modelo Poisson(λ): $P(k) = e^{-\lambda}\lambda^k/k!$
(2) modelo Exponencial(α): $P(k) = C_1 e^{-\alpha k}$
(3) modelo livre de escala: $P(k) = C_2 k^{-\gamma}$

onde C_1, $C_2 > 0$ são constantes (independentes de k). Suponha que determina rede fornece a seguinte amostra k_1, k_2, ..., k_n (assuma que a rede esteja constituída por n vértices).

(i) Obtenha o estimador de máxima verossimilhanção para λ, supondo que a rede possa ser descrita pelo modelo Poisson. (ii) O script do R `PoissonLik.R` simula uma amostra Poison(λ), grafica a verossimimlhança em função de λ, e mostra a estimativa para $\widehat{\lambda}_{MV}$ baseada na amostra. Carregue o script digitando `source(*/PoissonLik.R)` várias vezes e observe o resultado de cada vez (* é o path usual onde estão todos os scripts do curso). Mude o script considerando diversos valores para λ e execute o script. O valor de $\widehat{\lambda}_{MV}$ acompanha o valor de λ?

2.4 Projeto 2: um estimador para π

Georges-Louis Leclerc (1707-1788), Conde de Buffon, mostrou que vários problemas de probabilidade podem ser abordados utilizando argumentos de caráter geométrico. Em, particular, o problema conhecido hoje em dia como a agulha de Buffon permite realizar um experimento para estimar o valor de π.

Suponhamos que sobre um tabuleiro desenhamos linhas paralelas a distância t uma da outra. Posteriormente jogamos uma agulha de comprimento $l < t$ e observamos se esta cai ou não sobre alguma das linhas do tabuleiro. Surge assim naturalmente a seguintre pergunta: qual é a probabilidade de que a agulha esteja sobre uma linha t? Para respondermos esta questão, podemos parameterizar o espaço amostral (as posições das agulhas) da seguinte maneira. Seja Θ o angulo formado pela agulha e o conjunto de linhas t, e $X = (X_1, X_2)$ a posição do centro da agulha sobre o tabuleiro. Claramente, se ocorre o evento $\{X(\omega) \leq (l/2)\ \mathrm{sen}(\Theta(\omega))\}$, então a agulha corta uma linha t[2]. Não é difícil determinar a probabilidade deste evento pois as variáveis X e Θ são independentes e apresentam densidades uniformes nos intervalos $[0, t/2]$ e $[0, \pi/2]$ respectivamente,

$$f_X(x) = \begin{cases} 1/(t/2), & \text{se } 0 \leq x \leq t/2 \\ 0, & \text{caso contrário} \end{cases} \qquad f_\Theta(\theta) = \begin{cases} 1/(\pi/2), & \text{se } 0 \leq \theta \leq \pi/2 \\ 0, & \text{caso contrário} \end{cases}$$

Portanto a densidade conjunta do vetor (X, Θ) é simplesmente

$$f_{X,\Theta}(x, \theta) = \frac{4}{t\pi} \quad \text{quando} \quad (x, \theta) \in [0, t/2] \times [0, \pi/2],$$

e 0 no caso contrário. Logo

$$p = P\Big(X \leq \frac{l}{2}\ \mathrm{sen}(\Theta)\Big) = \int_0^{\pi/2} \int_0^{(l/2)\mathrm{sen}(\theta)} \frac{4}{t\pi}\, dx d\theta = \int_0^{\pi/2} \frac{4}{t\pi}\frac{l}{2}\mathrm{sen}(\theta) d\theta = \frac{2l}{t\pi}. \quad (3)$$

A formula (3) fornece indiretamente um estimador para π. De fato, se conseguimos uma estimativa para a probabilidade p, então (3) mostra como estimar $1/\pi$. Para simplificar a notação, seja E o evento $\{X \leq (l/2)\ \mathrm{sen}(\Theta)\}$, e logo seja $\xi(\omega) = \mathbf{1}_E(\omega)$, uma variável aleatória a qual é igual a 1 se a agulha touca a linha t e 0 no caso contrário: ξ é Bernoulli com probabilidade de sucesso $p = 2l/(t\pi)$. Seja $\xi_1, \xi_2, \ldots, \xi_n$, uma amostra desta população. No contexto da aplicação atual, esta amostra é interpretada como o resultado de jogar a agulha sobre o tabuleiro n vezes. Seguindo o procedimento agora ussual, utilizamos esta amostra para propor o estimador $\widehat{p} = \sum_{i=1}^n \xi_i/n$ para p. Desta maneira, de acordo com (3), podemos agora considerar o seguinte estimador para $1/\pi$

$$\widehat{\pi}^{-1} = \frac{t}{2l}\widehat{p}. \quad (4)$$

Exercício 23. (i) Qual é a distribuição da variável aleatória $\sum_{i=1}^n \xi_i$? (ii) Determine $\mathbb{E}[\sum_{i=1}^n \xi_i]$ e $\mathrm{Var}(\sum_{i=1}^n \xi_i)$. (iii) Calcule $\mathbb{E}[\widehat{p}]$ e $\mathrm{Var}(\widehat{p})$.

[2] faça um desenho!

Exercício 24. (i) Mostre que o estimador em (4) converge em probabilidade para π^{-1}, ou seja, que $\hat{\pi}^{-1}$ é (fracamente) consistente para π^{-1}.

Exercício 25. (i) Mostre que o estimador em (4) é não viciado, (ii) logo mostre que o EQM deste estimador é igual a

$$\frac{t\pi - 2l}{2ln\pi}$$

Desta última expressão podemos ver que o estimador em (4) é mais eficiente a medida que aumenta o comprimento da augulha l. Faça um gráfico do EQM em função de l, com l variando de 1 até 2.

Observação. O estimador para $1/\pi$ estudado neste projeto sugere o seguinte estimador para π,

$$\widehat{\pi} = \frac{2l}{t}\frac{1}{\widehat{p}}. \tag{5}$$

Este estimador é viciado para π mas neste caso é relativamente difícil determinar o vicio pois isto envolve calcular $\mathbb{E}\left[(\sum_i \xi_i)^{-1}\right]$, sendo $\sum_i \xi_i$ Binomial. Mesmo assim, é relativamente simples ver que $\hat{\pi}$ é consistente. Para isto ultimo é suficiente utilizar o mesmo argumento empregado no Exercício 10(v).

Exercício 26 (⮐)**.** Inicie R e carregue o código em `Buffon.R` fazendo

```
source("http://dcm.ffclrp.usp.br/~rrosales/aulas/Buffon.R")
```

Este script fornece quatro funções, `drawBuffon`, `runavrg`, `investPi` e `estPi`. `drawBuffon` mostra uma simulação do experimeto que consiste em jogar a agulha repetidas vezes (veja a figura 1), `runavrg` grafica uma estimativa para π conforme aumenta o número de vezes que é lançada a agulha (veja a figura 1). `investPi(N, l, t)` e `estPi(N, l, t)` fornecem respectivamente uma estimativa de $1/\pi$ e de π, sendo `N` o número de lançamentos da agulha, `l` é o comprimento da agulha e `t` a separação das linhas t. Estes parâmetros são inicializados para os valores `N=100`, `l=1`, e `t=2`, mas você pode mudar qualquer um a vontade (porém $l < t$). Utilize `investPi` para estudar as propriedades do estimador de $1/\pi$ com os seguintes valores de l: 0.5, 1 e 1.5. Digite, por exemplo,

```
y <- c();
for (i in 1:2000) y[i] <- investPi(N=2000);
```

Utilize as funções `var`, `mean` em `y` para verificar as conclusões obtidas analiticamente nos exercícios anteriores deste projeto.

2.5 Distribuições amostrais

Exercício 27. Uma variável de Bernoulli com probabilidade de sucesso p é amostrada, de forma, independente, duas vezes. Determine a função de probabilidade da média amostral.

Exercício 28. A variável aleatória ξ assome os valores $\{-2, -1, 1, 2\}$, cada um com a mesma probabilidade. Para uma amostra de tamanho dois, obtenha a distribuição de S^2 e verifique se ele é não viesado para estimar a variância de ξ.

Exercício 29. Coleta-se uma amostra de 10 observações independentes de uma população normal com média 2 e variância 2. Determine a probabilidade de a média amostral: (i) ser inferior a 1; (ii) ser superior a 2,5; (iii) estar entre 0 e 2.

Exercício 30. Um fabricante afirma que sua vacina contra COVID-19 imuniza em 90% dos casos. Uma amostra de 100 indivíduos que tomaram a vacina foi sorteada e testes foram feitos para verificar a imunização ou não desses indivíduos. (i) Se o fabricante estiver correto, qual é a probabilidade da proporção de imunizados na mostra ser inferior à 0,75? E superior à 0,9? (ii) Explique cuidadosamente quais argumentos desenvolvidos em sala de aula foram utilizados para calcular as probabilidades do item (i).

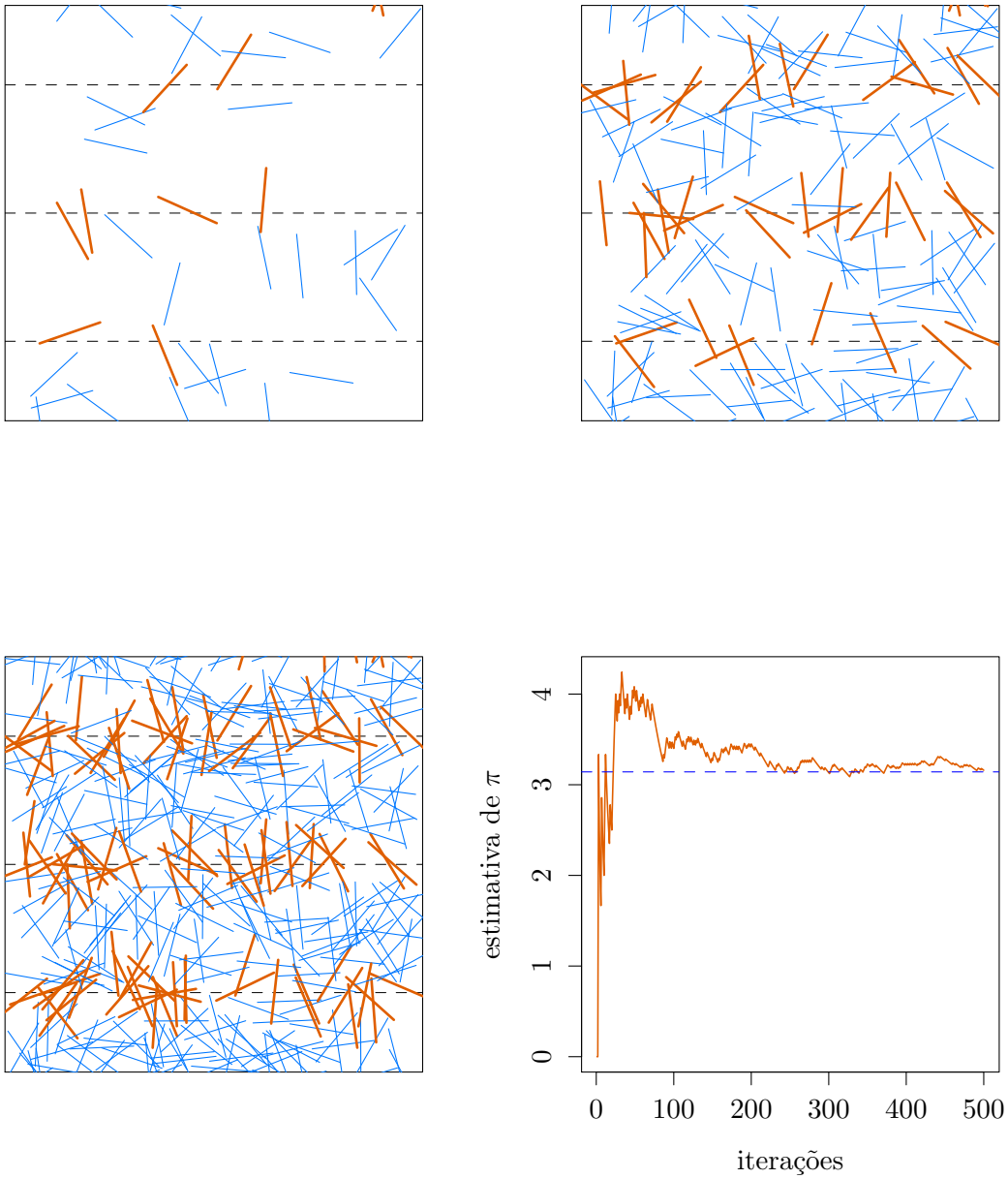

Figura 1: As três primeiras figuras mostram diversas simulações do experimento da agulha de Buffon para 60, 250, e 600 lançamentos da agulha. As agulhas que toucam uma banda t são mostradas em laranja. Estas figuras foram geradas com `drawBuffon`. A figura no canto inferior direito apresenta a convergência de uma estimativa para π gerada com `runavrg`.

Exercício 31. Este exercício fornece uma demonstração prática do Teorema Central do Limite para somas de variáveis aleatórias Bernoulli independentes e identicamente distribuídas.

Carregue o código `moedaCLT.R`, digitando desde o console do R[3]

```
source("http://dcm.ffclrp.usp.br/~rrosales/aulas/moedaCLT.R")
```

O código fornece a função `moedaCLT`, a qual pode ser utilizada para gerar m amostras independentes de n variáveis aleatórias Bernoulli(p) independentes. Pode pensar que esta função simula o lançamento de uma moeda n vezes e repite isto m vezes. `moedaCLT` a aceita três argumentos `N`, `M` e `p`: `N` corresponde a n, `M` corresponde a m e `p` a p, a probabilidade de sair cara em qualquer lançamento, e retorna o vetor

$$\left(\frac{S_n^1}{n}, \frac{S_n^2}{n}, \ldots, \frac{S_n^m}{n}\right),$$

onde S_n^i/n, $i = 1, \ldots, m$, corresponde a proporção de caras após de jogar a moeda n vezes no i-ésimo experimento. Por exemplo,

```
v1 <- moedaCLT(N=10000, M=30000, p=0.5);
```

simula o lançamento de uma moeda (honesta) 10000 vezes, repete isto 30000 vezes calculando de cada vez a fração relativa de caras, e finalmente guarda estes valores no vetor `v1`. Digite

```
hist(v1,breaks=60, main="", ylab="frequencia",xlab="Zn")
```

A função `hist()` calcula o histograma de `v1`, isto é $\widehat{h}_{S_n^1/n,\ldots,S_n^m/n}$, e apresenta o grafico desta função. Utilice várias vezes `moedaCLT()` tentando valores diferentes para `M` e `N` de cada vez. (i) Consegue enxergar o Teorema Central do Limite? (ii) Qual dos argumentos `N` ou `M` controla a convergência no Teorema Central do Limite? qual controla a convergência do histograma em (2)?

Exercício 32. (Continuação do **Exercício 26**.) Inicie R e digite

```
source("http://dcm.ffclrp.usp.br/~rrosales/aulas/Buffon.R")
y <- c(); for (i in 1:2000) y[i] <- estPi(N=2000);
hist((y-mean(y))/sd(y), breaks=50, col="lightblue")
lines(seq(-3,3,0.01),dnorm(seq(-3,3,0.01)), lwd=2)
```

A segunda linha gera 2000 estimativas para π, de acordo ao estimador em (5), guardando-as no vetor `y`. Cada estimativa é obtida ao simular o lançamento da agulha 2000 vezes. O histograma gerado na terceira linha do código acima com `hist` sugere que a distribuiçõ amostral de $\hat{\pi}$ é normal. A ultima linha grafica a densidade normal padrão. Repita esta analise variando de cada vez o valor de `N` em `estPi`, utilizando por exemplo os valores `100`, `200`, `500` e `5000`.

Exercício 33. (Continuação do **Exercício 22**.) Utilice o script `PoissonLik.R` para gerar valores de $\widehat{\lambda}_{MV}$, fixando por exemplo o valor de λ em 30. (i) Faça um histograma para os valores de $\widehat{\lambda}_{MV}$. (ii) Explique por que a distribuição amostral de $\widehat{\lambda}_{MV}$ deve convergir a uma distribuição normal.

Exercício 34. Desejamos coletar uma amostra de uma variável aleatória X com distribuição normal de média desconhecida e variância 30. Qual deve ser o tamanho da amostra para que, com 0,92 de probabilidade, a média amostral não difira da média da população por mais de 3 unidades?

[3] alternativamente pode baixar este arquivo no seu micro para carrega-lho posteriormente como

```
source("C://lugar_do_download_no_seu_micro//moedaCLT.R")
```

assumendo que você trabalha em Windows. Caso você esteja trabalhando em Linux (ou numa Mac) troque o delimitador de pastas "`//`" por "`/`".

3 Intervalos e testes de hipótese

3.1 Intervalos de Confiança

Exercício 35. Por analogía a produtos similares, o tempo de reação de um novo medicamento pode ser considerado como tendo distribuição normal com média μ e variaância 4. Vinte pacientes foram sorteados, receberam o medicamento e tiveram seu tempo de reação anotado. Os dados foram os seguintes: 2,9; 3,4; 3,5; 4,1; 4,6; 4,7; 4,5; 3,8; 5,3; 4,9; 4,8; 5,7; 5,8; 5,0; 3,4; 5,9; 6,3; 4,6; 5,5 e 6,2. Obtenha intervalos de confiança para o tempo médio de reação para: (i) γ=96%, (ii) γ=75%.

Exercício 36. Uma amostra de 25 observações de uma normal com média μ desconhecida e variância 16 foi coletada e forneceu uma média amostral de 8. Construa intervalos com confiança 80%, 85%, 90% e 95% para a média populacional. Comente as diferenças encontradas.

Exercício 37. Será coletada uma amostra de uma população normal com desvio padrão igual a 9. Para uma confiança de γ=90%, determine a amplitude do intervalo de confiança para a média populacional nos casos em que o tamanho da amostra é 30, 50 ou 100. Comente as diferenças.

Exercício 38. Numa pesquisa com 50 eleitores, o candidato A obteve 0,42 da preferência dos eleitores. Construa, para a confiança 95%, os intervalos otimista e conservador de confiança para a proporção de votos a serem recebidos pelo candidato mencionado, supondo que a eleição fosse nesse momento.

Exercício 39. Interprete e comente as afirmações: (i) A média de salário inicial para recém formados em informática biomédica está entre 7 e 9 salários mínimos com confiança 95%. (ii) Quanto maior for o tamanho da amostra, maior é a probabilidade da média amostral estar próxima da verdadeira média.

Exercício 40. O intervalo [35,21; 35,99], com confiança 95% foi construído a partir de uma amostra de tamanho 100, para a média μ de uma população normal com desvio padrão igual a 2. (i) Qual e o valor encontrado para a média dessa amostra? (ii) Se utilizássemos essa mesma amostra, mas uma confiança de 90%, qual seria o novo intervalo de confiança?

Exercício 41. Uma industria farmacêutica tem interesse em determinar a probabilidade p da efetividade de um novo fármaco. Uma amostra piloto de tamanho 100 revelou que 60% dos pacientes submetidos ao tratamento com o fármaco tiveram resultados favoráveis. (i) Utilizando a informação da amostra piloto, determine o tamanho da amostra para que, com 0.8 de probabilidade, o erro cometido na estimação seja no máximo 0.05. (ii) Se na amostra final, com tamanho obtido em (i), observou-se que 51% dos pacientes tiveram resultados favoráveis, construa um intervalo de confiança para p, com confiança 95%.

Exercício 42. (Intervalo para $\mu_1 - \mu_2$) O arquivo
http://dcm.ffclrp.usp.br/~rrosales/aulas/trabalho.txt
apresenta os dados referentes a taxa de trabalho infantil em Brasil para criaças de diferentes raças durante o período 1992-2008[4]. A taxa de trabalho infantil é definida como o percentual da população residente de 10 a 15 anos de idade que se encontra trabalhando ou procurando trabalho na semana de referência, em determinado espaço geográfico, no ano considerado. (i) Construa um intervalo de confiança de 95% para a diferença entre as taxas de trabalho médias para crianças brancas e pretas. (ii) Interprete o intervalo obtido em (i), isto é, qual é o

[4]Fonte: Instituto Brasileiro de Geografia e Estatística (IBGE). Série: CAJ421 - Taxa de trabalho infantil, por cor
http://seriesestatisticas.ibge.gov.br/series.aspx?vcodigo=CAJ421

valor p	interpretação
$p < 0.01$	evidência forte contra H_0
$0.01 \leq p < 0.05$	evidência moderada contra H_0
$0.05 \leq p < 0.10$	evidência fraca contra H_0
$0.10 \leq p$	não a evidência contra H_0

Tabela 1: interpretação do p-valor

significado deste intervalo? (iii) Quais são os supostos necessários para construir o intervalo? (iv) Você acredita que os supostos são satisfeitos neste caso? (v) Construa um intervalo comparando as crianças brancas e indígenas. Interprete os seus resultados.

Exercício 43. O seguinte exercício tem como objetivo ilustrar a interpretação ussual de um intervalo de confiança. Gere uma amostra de tamanho 20 da distribuição normal com média 0 e desvio padrão 5 (por exemplo utilizando R). Calcule o intervalo de confiança para a média baseado na amostra com coeficiente $\gamma = 0,95$, por exemplo. Repeta estes passos 100 vezes e conte o número de vezes nas quais o intervalo captura o verdadeiro valor de μ (a média populacional). Divida esta frequencia pelo número total de repetições e compare o valor final com γ. Sugestão: utilice as funções `rnorm`, `qnorm`, `mean`.

Exercício 44. Construa um intervalo de confiança para o parâmetro λ de uma populacão Poisson. Suponha que o tamanho das amostras a serem utilizadas seja suficientemente grande; por exemplo $n \geq 30$.

3.2 Testes de Hipóteses

Nível descritivo (p-valor)

Em lugar de fixar o nível de um teste de hipótese, R e outros pacotes fornecem uma quantidade conhecida como o p-valor do teste. Este ultimo pode ser utilizado para rejeitar ou não a hipótese nula. Suponhamos que o estatístico $\widehat{\theta}$ é considerado em um teste para o parâmetro θ. Seja $\widehat{\theta}(x)$ a estimativa de $\widehat{\theta}$ baseada nos valores da amostra $x = (x_1, x_2, \ldots, x_n)$. Suponhamos que ao fixamos o nível α definimos a região crítica $\mathcal{R}$, e assim optamos pela rejeição de H_0 sempre e quando $\widehat{\theta}(x) \in \mathcal{R}$. Alternativamente, em lugar de fixar o nível α, podemos calcular a probabilidade

$$p = \mathbb{P}\big(\{\omega : \widehat{\theta}(\omega) \geq \widehat{\theta}(x)\}|H_0\big), \tag{6}$$

e rejeitar a hipótese nula quando o valor de p for pequeno, por exemplo $p < \alpha$, onde α típicamete determina o nivel do teste. A probabilidade p calculada em (6), utilizada para rejeitarmos ou não H_0, é conhecida como o p-valor do teste. Usualmente, o valor p é utilizado seguindo os criterios apresentados na Tabela 1.

Destacamos que o p-valor de um teste realmente é a variável aleatória $p : \Omega \to \mathbb{R}$ definida por

$$\tilde{\omega} \mapsto \mathbb{P}\Big(\big\{\omega : \widehat{\theta}(\omega) \geq \widehat{\theta}\big(X(\tilde{\omega})\big)\big\}\big|H_0\Big), \qquad \tilde{\omega} \in \Omega,$$

onde $X(\tilde{\omega}) = (X_1(\tilde{\omega}), \ldots X_n(\tilde{\omega}))$ é a amostra. Isto ultimo é importante quando são estudadas as propriedades de um p-valor, mas não faremos referência a isto durante o curso.

3.2.1 Testes para μ e p (uma ou duas populações)

Exercício 45. Uma variável aleatória tem distribuição normal e desvio padrão igual a 12. Estamos testando se sua média é igual ou é diferente de 20 e coletamos uma amostra de 100

valores dessa variável, obtendo uma média amostral de 17,4. (i) Formule as hipóteses. (ii) Obtenha a região crítica e dê a conclusão do teste para os seguintes níveis de significância: 1%, 2%, 4%, 6% e 8%.

Exercício 46. Para uma variável aleatória com densidade normal e desvio padrão 5, o teste da média μ=10 contra μ=14, teve a região crítica dada por $\{x \in \mathbb{R} : x > 12\}$ para uma amostra de tamanho 25. Determine as probabilidades dos erros tipo I e II.

Exercício 47. Uma máquina deve produzir peças com diâmetro de 2 cm. Entretanto, variações acontecem e vamos assumir que o diâmetro dessas peças siga o modelo Normal com variância igual a 0,09 cm^2. Para testar se a máquina está bem regulada, uma amostra de 100 peças é coletada. (i) Formule o problema como um teste de hipóteses. (ii) Qual seria a região crítica se $\alpha = 0,02$? (iii) se a região de aceitação fosse $\{x \in \mathbb{R} | 1,95 \leqslant x \leqslant 2,05\}$, qual seria o nível de significância do teste? Nesse caso, determine a probabilidade do erro tipo II se μ =1,95 cm. (iv) Se para essa amostra $\bar{x} = 1,94$; qual a decisão em (ii)?, em (iii)?

Exercício 48. A vida média de uma amostra de 100 lâmpadas de certa marca é 1615 horas. Por similaridade com outros processos de fabricação, supomos o desvio padrão igual a 120 horas. Utilizando α=5%, desejamos testar se a duração média de todas as lâmpadas dessa marca é igual ou é diferente de 1600 horas. Qual é a conclusão? Determine também a probabilidade do erro tipo II, se a média fosse 1620 horas.

Exercício 49. Em certo estudo foi considerada a influência da succinilcolina nos níveis de circulação de andrógeno no sangue. Amostras de sangue de diversos indivíduos foram obtidas antes e depois de 30 minitos da injeção de succinilcolina. Os níveis de andrógeno para etas condições são apresentados abaixo.

antes	5,18	3,05	4,10	7,05	2,68
depois	3,10	3,99	5,21	10,26	5,44

(i) Assumindo que as populações de andrógeno antes e depois de 30 minutos da injeção são normais, teste ao nível de 5% se as concentrações de andrógeno são alteradas. (ii) Determine o p-valor do teste.

Exercício 50. Um criador tem constatado uma proporção de 10% do rebanho com verminose. O veterinário alterou a dieta dos animais e acredita que a doença diminuiu de intensidade. Um exame em 100 cabeças do rebanho, escolhidas ao acaso, indicou 8 delas com verminose. Ao nível de 8%, há indícios de que a proporção diminuiu?

Exercício 51. Considere o teste $p = 0,6$ contra $p \neq 0,6$. Sendo $n = 100$, indique a probabilidade de erro tipo I para as seguintes regiões críticas: (i) $\mathcal{R} = \{x \in \mathbb{R} | x < 0,56 \text{ ou } x > 0,64\}$, (ii) $\mathcal{R} = \{x \in \mathbb{R} | x < 0,54 \text{ ou } x > 0,66\}$.

Exercício 52. Durante uma epidemia de resfriado, 2.000 crianças foram pesquisadas por uma renomada indústria farmacêutica para determinar se o novo medicamento da empresa é eficaz após dois dias. Entre 120 crianças que tiveram resfriado e receberam o medicamento, 29 foram curadas dentro desse prazo. Entre 280 crianças que não receberam o medicamento, 56 foram curadas em dois dias. Há alguma indicação significativa que apóia a afirnação da empresa sobre a eficácia do medicamento?

Exercício 53. Um estudo foi realizado para determinar se diferenças no nível do anticorpo IgG (Imunoglobulina G) afeita o desenvolvimento de trombose. A seguinte tabela resome os dados.

Grupo	nível médio de IgG (ml/u)	n	desvio
com Trombose	59.01	23	44.89
sem Trombose	46.61	24	34.85

(i) Qual é a conclussão ao nível de 1%? (ii) Quais supostos foram feitos para realizar o teste?

Exercício 54. Acredita-se que a probabilidade de que uma criança recem nascida seja menina é maior da probabilidade de ser menino. Com o objetivo de verificar esta hipótese foi considerada uma amostra de 25468 crianças das quais 13173 resultaram ser meninas. Estes dados confirmam o suposto feito inicialmente?

Exercício 55. Um geneticista esta estudando a proporção de homens e mulheres com determinado distúrbio sanguíneo. Em uma amostra aleatória de 100 homens, 31 são afetados, enquanto apenas 24 das 100 mulheres testadas apresentam o distúrbio. Podemos concluir ao nível de 0.01 que a proporção de homens afetados pelo distúrbio é significativamente maior do que a proporção de mulheres afligidas?

Exercício 56. Estudos preliminares sugerem que a proporção de bases mutadas em determinada região do DNA é 0.6. Para testar essa hipótese, uma amostra aleatória de 200 sequencias da região em questão foi selecionada. Se o número de mutações estiver entre 110 e 130, não devemos rejeitar os resultados de estudos preliminares. (i) Determine o nível do teste. (ii) Determine o poder se a proporção alternativa é 0.5 ou 0.7. (iii) Em base aos resultados em (i) e (ii), este procedimento corresponde a um bom teste?

3.2.2 *t-Student*: testes e intervalos para μ com σ^2 desconhecida

Exercício 57. Com auxílio da tabela *t-Student* calcule (se necessário, aproxime):
(i) $P(-3,365 \leqslant t_5 \leqslant 3.365)$. (ii) $P(|t_8| < 1.4)$. (iii) $P(-1,1 \leqslant t_{14} < 2.15)$. (iv) $a : P(t_9 > a) = 0.02$. (v) $b : P(t_{16} \leqslant b) = 0.05$. (vi) $c : P(|t_{11}| \leqslant c) = 0.1$. (vii) $d : P(|t_{21}| > d) = 0.05$.

Exercício 58. Uma amostra de 20 observações de uma variável com distribuição normal foi colhida, obtendo-se desvio padrão 1,1. No teste μ=5 contra $\mu > 5$, foi estabelecida a região critica $\{t \in \mathbb{R} | t > 2,033\}$. Determine a probabilidade do erro tipo I.

Exercício 59. A porcentagem anual média da receita municipal empregada em saneamento básico em pequenos municípios de um estado tem sido 8% (admita que esse índice se comporte segundo um modelo normal). O governo pretende melhorar esse índice e, para isso, ofereceu alguns incentivos. Para verificar a eficácia dessa atitude, sorteu 10 cidades e observou as porcentagens 8, 12, 16, 9, 11 e 12. Os dados trazem evidência de melhoria, ao nível de 2%? Caso altere a média, dê um intervalo de confiança para anova média.

Exercício 60. Um pesquisador deseja estimar o nível de expressão média de um determinado gene. Uma amostra de 5 observações fornece os seguintes níveis de expressão: 75, 85, 95, 105 e 115. (i) Supondo que a expressão do gene seja uma variável aleatória normalmente distribuída, determine o intervalo de confiança de 95% para o nível médio da expressão do gene.

Exercício 61. Os seguintes dados correspondem a pressão arterial sistólica (mmHg) de 12 pacientes submetidos à terapia medicamentosa para hipertensão,

183 152 178 157 194 163 144 114 178 152 118 158

(i) Podemos concluir com base nesses dados que a média populacional é menor que 165? Considere um nível de significância de 5%. (ii) Quais suposições relativas a amostra são necessárias para realizar o teste?

Exercício 62. As porcentagens de α_2-Macroglobulina de 5 pessoas com baixo peso e 4 obesas aparecem na tabela a seguir.

Baixo Peso	6,13	7,05	7,48,	7,53	8,40
Obeso	8,79	9,19	9,21	9,97	

Verifique se o nível médio desta proteína é significativamente diferente nos dois grupos de pessoas consideradas. Considere um nível de significância de 5%.

Exercício 63. Cinco sujeitos foram considerados em um experimento para determinar se a exposição ao monóxido de carbono afeita a capacidade respiratória. Os sujeitos foram expostos a câmaras respiratórias, uma sem CO e a outra com uma alta concentração de CO. As medidas de frequência respiratória foram feitas para cada sujeito para cada câmara. Os resultados são apresentados pela seguinte tabela.

Indivíduo	Com CO	Sem CO
1	30	30
2	45	40
3	26	25
4	25	23
5	34	30

Teste ao nível $\alpha = 0,05$ se existe diferênça entre os dois tratamentos. Suponha que a frequência respiratória seja aproximadamente normal.

Exercício 64. Inicie R e carregue os dados `energy.txt` no site do curso digitando

```
dt <- read.table(file="http://dcm.ffclrp.usp.br/~rrosales/aulas/energy.txt",
head=TRUE)
attach(dt)
```

Estes dados contém duas colunas: `expend` e `stature`, e representam o consumo energetico de mulheres magras (lean) e obesas (obese). O argumento `head=TRUE` da função `read.table` permite Digite

```
t.test(expend~stature, paired=TRUE)
```

A função `t.test`, com a sintaxe acima, permite realizar um teste t utilizando o estimador

$$T = (\bar{X}_2 - \bar{X}_1) \Big/ \sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{n}}$$

(i) No caso dos dados em `energy.txt`, quais são as hipóteses H_0 e H_a que estão sendo testadas? (ii) Qual é o resultado do teste? (iii) A figura 2 mostra a função poder para o teste em (i), para três valores de α: 0.001, 0.01 e 0.05. A figura mostra que o poder do teste para $\alpha = 0.05$ é maior ao poder dos testes para os níveis 0.001 e 0.01, por que? (iii) Escreva um código em R, o qual permita calcular a função poder para testes t-Student. (Sugestão: utilice a função `qt`.)

Exercício 65. Suponha que o tempo de vida T de um paciente submetido a uma determinada intervenção possua distribuição exponencial com densidade

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0.$$

É considerada a amostra $T_1, T_2, \ldots, T_n$ constituída pelos tempos de sobrevida de n pacientes com o objetivo de testar a hipótese nula $H_0 : \lambda = 2$ versus a hipótese alternativa $H_a : \lambda > 2$. Seja W o menor valor da amostra e Z o maior valor da amostra. É considerado um teste com região crítica da forma $W < k$. (i) Determine o valor de k em termos de n, de tal forma que o teste possua um nível de significância de 5%. (ii) Obtenha uma expressão para a função poder do teste. (iii) Considere as questões (i) e (iii) agora para uma região crítica da forma $Z < k$. (iv) Qual dos dois testes possue maior poder? Faça um gráfico de ambas funções poder com $n = 10$.

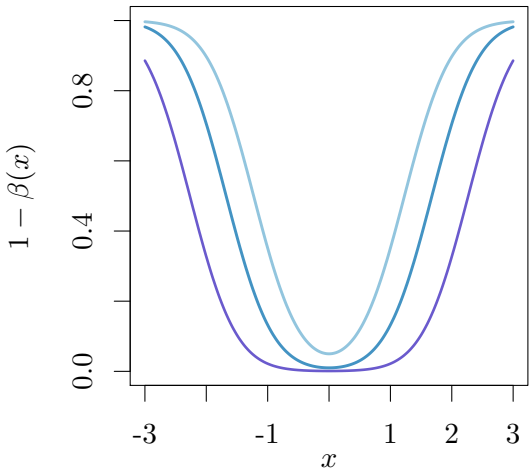

Figura 2: funções poder para o teste do Exercicio 64, para três níveis α: 0.001, 0.01 e 0.05.

Exercício 66. Carregue os dados `chiken.txt`. Estes dados contem o efeito de duas dietas diferentes no crecimento de frangos durante as primeiras semanas de vida. Os dados apresentam quatro colunas: "`weight`", "`Time`", "`Chick`", e "`Diet`". A figura 3 apresenta um "Box Plot"[5] gerado utilisando a sintaxe `boxplot(weight~Diet)`. (i) Em base ao gráfico, diga se os dois tratamentos tem algum efeito sobre o peso médio dos frangos. (ii) Faça um teste de hipotese para verificar a sua opinião. Qual é a sua conclusão? [Sugestão: veja o exercício anterior]

Exercício 67. Inicie R e carregue os dados `trabalho.txt`. Este arquivo contém os dados do Exercicio 42. (i) Faça um teste para verificar se no Brasil existe diferença na taxa de trabalho de crianças pretas e crianças brancas. Qual é a sua conclusão? (ii) Os resultados aqui são consistentes com aqueles obtidos no Exercicio 42?

3.2.3 χ^2: testes e intervalos para a Variância

Exercício 68. Seja χ^2 uma variável aleatória com distribuição qi-quadrado. Considere a tabela χ^2 e o valor de x em cada um dos seguintes casos:

$$(i).\quad P(\chi^2 < x) = 0.05,\ gl = 7, \qquad (ii).\quad P(\chi^2 \geq x) = 0,1,\ gl = 16,$$
$$(iii).\quad P(|\chi^2| > x) = 0.01,\ gl = 10, \qquad (iv).\quad P(|\chi^2| \leq x) = 0.5,\ gl = 8.$$

Exercício 69. O tempo de certo evento observado em 18 provas forneceu a estimativa para S de 6,3 (ns). Obtenha um intervalo de confiança de 95% para a variância da população, σ^2. Suponha que a distribuição dos tempos observados seja normal.

[5] A barra inferior representa a menor observação não extrema, o borde inferior da caixa corresponde ao primeiro quartil Q_1 (i.e. o valor de x tal que $\widehat{F}_{x_1,\ldots,x_n}(x) = 0,25$), a barra cheia é a mediana dos dados, o borde superior da caixa é o terceiro quartil $Q_3 = x : \widehat{F}_{x_1,\ldots,x_n}(x) = 0,75$, e a barra superior representa a maior observação não extrema. Os símbolos $\circ$ representam eventos moderadamente extremos. Um dado é considerado moderadamente extremo se o seu valor esta entre $1,5(Q_3 - Q_1)$ e $3(Q_3 - Q_1)$. Se o valor de uma observação é maior do que $3(Q_3 - Q1)$, então esta é representada com o símbolo $*$ e considerado como um verdadeiro extremo.

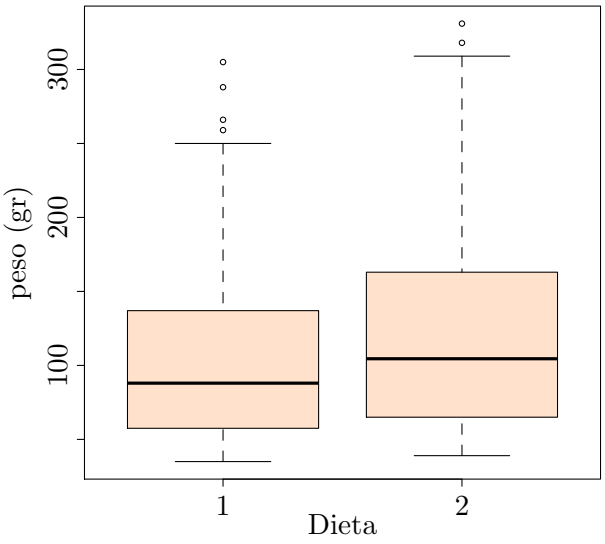

Figura 3: Box Plots para os dados em `chiken.txt`.

3.2.4 Teste F (Fisher-Snedecor): σ_1^2/σ_2^2

Exercício 70. Supondo $X \sim F(a, b)$, encontre x_c tal que: (i) $P(X \geqslant x_c) = 0,05$ com a=18, b=3. (ii) $P(X > x_c) = 0,05$ com a=3, b=18. (iii) $P(X > x_c) = 0,05$ com a=180, b=192. (iv) $P(X \geqslant x_c) = 0,95$ com a=5, b=12. (v) $P(X \geqslant x_c) = 0,95$ com a=30, b=40.

Exercício 71. Queremos comparar três hospitais, a través da satisfação demonstrada por pacientes quanto ao atendimento, durante o período de internação. Para tanto, foram selecionados, aleatoriamente, pacientes com grau de enfermidade semelhante. Cada paciente preencheu um questionário e as respostas geraram índices variando de 0 a 100, indicando o grau de satisfação. Os resultados foram

	Hospital		
	A	B	C
n	10	15	13
$\bar{x}$	80,7	59,0	72,3
$s^2(x)$	113,3	101,4	106,5

(i) Baseando-se nos dados apresentados, teste a igualdade das variâncias para os hospitais A e B. Use $\alpha = 0.10$. (ii) Teste se as médias populacionais são iguais. Utilice $\alpha = 5\%$. Qual é a sua conclusão?

Exercício 72. Dois tipos de instrumentos são utilizados para medir a quantidade de monóxido sulfúrico na atmosfera devem ser comparados. Dese-jamos determinar se os dois intrumentos rendem medições com a mesma variabilidade. As seguintes leituras foram registradas.

monóxido sulfúrico	
Intrumento A	Instrumento B
0,86	0,87
0,82	0,74
0,75	0,63

Considere o teste $H_0 : \sigma_A = \sigma_B$, $H_a : \sigma_A \neq \sigma_B$, e utilice o seguinte resultado: Se F_α é Fisher(n, m), então $F_{1-\alpha}(m, n) = 1/F_\alpha(n, m)$.

Exercício 73. Procure e carregue os dados `stroke.txt`. Entre outras informações, estes dados fornecem a idade de pessoas de ambos sexos na Estônia as quais sofreram um enfarto durante o período 1991-1993. Digite `var.test(age~sex)`. (i) O que esta sendo testado (quais são as hipóteses?) (ii) Baseado no valor p do teste, qual é a sua conclusão?

3.3 Projeto 3: Bioinformática

O objetivo deste exercício é aplicar alguns dos métodos utilizados em seções anteriores a uma base de dados constituída pelos valores da expressão genica de pacientes com leucemia linfóide e leucemia mielóide aguda. Os dados a serem utilizados foram tomados do pacote `multtest`, o qual forma parte de Bioconductor: www.bioconductor.org, e estão baseados nas análises em [6]. Os dados podem ser carregados como

```
library(multtest); data(golub);
```

caso `multtest` esteja instalado, ou directamente do site do curso digitando

```
load(url("http://dcm.ffclrp.usp.br/~rrosales/aulas/r-data-stat-IBM/golub.RData"))
```

Os dados disponíveis na matriz `golub` apresentam os valores da expressão de 3051 genes (filas) de 38 pacientes diagnosticados com leucemia (colunas). Os dados dos primeiros 27 pacientes correspondem a pessoas com leucemia linfóide (ALL) e os ultimos 11 a pessoas com leucemia mielóide aguda (AML). O tipo do tumor se encontra indicado pelo vetor numerico `golub.cl`, onde a condição ALL é determinada pelo número `0` e AML pelo número `1`. Os nomes dos genes se encontram em `golub.gnames`, uma matrix com 3 colunas: um índice para o gene, a identidade do gene, e o nome do gene. Por exemplo, o gene `M92287_at` identificado com "CNND3 Cyclin D3" corresponde a file número 1042 em `golub.names`,

```
 golub.gnames[1042,]
[1] "2354"     "CCND3 Cyclin D3"     "M92287_at"
```

Assim,

```
golub[1042,2]
[1] 1.52405
```

representa a expressão do gene `M92287_at` para o paciente 2. `golub[,1]` representa os valores da expressão para os 3051 genes do paciente 1, e `golub[1024,]` os valores da expressão do gene `M92287_at` para todos os 38 pacientes,

```
 golub[1024,]
 [1] -1.45769 -1.39420 -1.46227 -1.40715 -1.42668 -1.21719 -1.37386 -1.36832
 [9] -1.47649 -1.21583 -1.28137 -1.03209 -1.36149 -1.39979 -1.39503 -1.40095
[17] -1.56783 -1.20466 -1.24482 -1.60767 -1.06221 -1.12665 -1.20963 -1.48332
[25] -1.25268 -1.27619 -1.23051 -1.43337 -1.08902  0.40633 -1.26183 -1.44434
[33] -1.47218 -1.34158 -1.22961 -0.39456 -1.34579 -1.32403
```

Suponhamos que desejamos separar os valores da expressão do gene `M92287_at` em dois grupos: ALL, AML (segundo o tipo de tumor). Definimos primeiro uma variável do tipo *factor* com nome `gol.fact`,

```
gol.fac <- factor(golub.cl, levels=0:1, labels=c("ALL", "AML"))
```

Agora, para obter os valores de expressão de `M92287_at` para os pacientes do grupo ALL fazemos

```
golub[1042, gol.fact=="ALL"]
```

Esta maneira de organizar os dados permite por exemplo calcular a expressão genica média (para cada gene) de tudos os pacientes do tipo ALL,

```
mediaALL <- apply(golub[, gol.fac=="ALL"], 1, mean)
```

(veja `help(apply)`). A média de cada um dos 3051 genes dos dados do tipo ALL se encontra no vetor `mediaALL`. Suponhamos agora que temos interesse em estudar o gene identificado por CD33 (segundo [6], este gene pode ser utilizado para identificar células do tipo linfóide das mielóides!). Para saber o indice da fila de `golub` para este gene fazemos

```
grep("CD33", golub.gnames[,2])
[1] 808
```

isto é, os valores da expressão para o antigeno CD33 se encontram em `golub[808, ]`.

Exercício 74. Digite

```
mall <- apply(golub[,gol.fac=="ALL"], 1, mean)
maml <- apply(golub[,gol.fac=="AML"], 1, mean)
o <- order(abs(mall-maml), decreasing=TRUE)
print(golub.gnames[o[1:5],2])
```

Interprete o resultado e diga qual é a sua importância.

Exercício 75. Utilice a função `grep` para encontrar os oncogenes em `golub`. (i) Quantos oncogenes tem a base de dados? (ii) Encontre os nomes dos oncogenes com o maior valor de expressão médio para os pacientes do tipo ALL. (iii) Faça o mesmo para os pacientes do tipo AML.

Exercício 76. Escolha os dados do gene CD33. (i) Faça um teste de hipotese para verificar a igualdade das variâncias na expressão do gene CD33 nos grupos ALL e AML. (ii) Considere um teste para verificar a igualdade no nível médio da expressão do gene CD33 nos grupos ALL e AML.

Exercício 77. O oncogene "MYBL2 V-myb avian myeloblastosis viral oncogene homolog-like 2" se encontra na fila 1788 de `golub`. (i) Utilice um boxplot para comparar os dois grupos ALL e AML. Você acredita que o nível de expressão médio varia de acordo com o grupo? (ii) Considere um t-teste para verificar se o valor médio de expressão é igual. (iii) Repita estes analises para o gene "HOXA9 Homeo box A9", o qual segundo [6] causa leucemia. Qual é a sua conslusão?

4 Análise de variância e regressão linear

Exercício 78. Três diferentes hospitais do mesmo porte em Ribeirão Preto desejam testar se estes apresentam movimento médio equivalente. Foi escolhida uma semana típica de trabalho e o desempenho nesses dias foi registrado. Os dados obtídos em unidades de dinheiro arbitrárias são apresentados na seguinte tabela

Banco		
1	2	3
146,4	194,3	173,7
199,2	227,2	246,5
179,5	203,4	289,8
98,4	111,8	127,4
263,7	275,0	265,6

Qual seria a sua conclusão ao nível α =5%?

Exercício 79. Um estudo deseja avaliar o efeito do treinamento no tempo de reação de atletas submetidos a um certo estímulo. O treinamento consiste na repetição de um movimento e foi utilizada uma amostra de 37 atletas. Para cada atleta foi atribuído um certo número de repetições X e, então, foi medido o tempo de reação Y, em milisegundos. Uma reta de mínimos quadrados foi ajustada aos dados, fornecendo a equação

$$\widehat{y}_i = 80,5 - 0,9x_i, \quad i = 1, \ldots, n.$$

(i) Qual é o significado das estimativas para α e β?

Exercício 80. Inicie R e carregue os dados `cabbages.txt`. Estes dados contem informações sobre plantios de repolhos e estão constituídos por quatro colunas: `Cult`: origem do cultivo, `Date`: data da plantação, `HeadWt`: peso da cabeça do repolho (em Kg), `VitC`: conteúdo de ácido ascorbico (vitamina C, em unidades arbitrárias). Ao digitar

```
minharegressao <- lm(HeadWt~VitC)
```

deverá aparecer

```
Call:
lm(formula = HeadW~VitC)

Coefficients:
(Intercept)         VitC
    5.92806     -0.05754
```

O argumento a `lm` é a *fórmula de um modelo*. Na sua forma mais simples, o modelo `y~x` indica que `y` é a variável dependente e `x` a variável independente (esta última é conhecida em uma regressão como a variável descritiva). Neste caso, como saídas de `lm` obtemos o intercepto (β) com o eixo y e a inclinação (α) da reta que melhor descreve os dados. A estimativa para a reta de regressão portanto é

$$\texttt{HeadWt} = 5.92806 - 0.05754 \times \texttt{VitC}.$$

Maiores informações sobre a regressão são obtidos ao escrever

```
summary(minharegressao)
```

o qual gera a seguinte informação

```
Call:
lm(formula = HeadWt ~ VitC)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0150 -0.5117 -0.1575  0.4244  1.6095

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.928059   0.505983  11.716  < 2e-16 ***
VitC        -0.057545   0.008603  -6.689 9.75e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6687 on 58 degrees of freedom
Multiple R-squared: 0.4355,        Adjusted R-squared: 0.4257
F-statistic: 44.74 on 1 and 58 DF,  p-value: 9.753e-09
```

`Residuals` fornece algumas propriedades que resumem a distribuição dos erros e_i. Lembramos que a distribuição de estes apresenta *a priori* média 0, portanto a mediana dos erros deve estar próxima de este valor (neste caso -0.1575). `Coefficients`; mostra novamente as estimativas para β e α e para cada uma o seu erro padrão, testes t, e p-valores. Os símbolos a direita correspondem a um indicador gráfico do nível do teste; `*` significa $0,01 < p < 0,05$ (veja a linha `Signif.codes:...`). `Residual standard error` é a variação residual, uma quantidade que mede a variabilidade das observações a respeito da reta de regressão, e fornece uma estimativa para σ, a variância dos e_i. `Multiple R-squared` é o coeficiente de correlação de Pearson. `F-statistics` corresponde ao resultado do teste H_0: $\alpha = 0$, H_a: $\alpha \neq 0$. Finalmente, os comandos

```
plot(VitC,HeadWt,xlab="concentracao de vitamina C (unidades
        arbitrarias)", ylab="peso da cabeca do repolho (Kg)",
        cex=0.9, lwd=0.65)
abline(lm(HeadWt~VitC), lwd=1.5, col="navy", lty=2)
```

produzem a figura 4. (i) Baseado em estes resultados, você acredita que o modelo de regressão linear é apropriado em este exemplo? Qual dos resultados fornecidos por R levo você a sua conclusão? (ii) Qual é o peso esperado de uma cabeça de repolho com 60 unidades de vitamina C? e para 100 unidades?

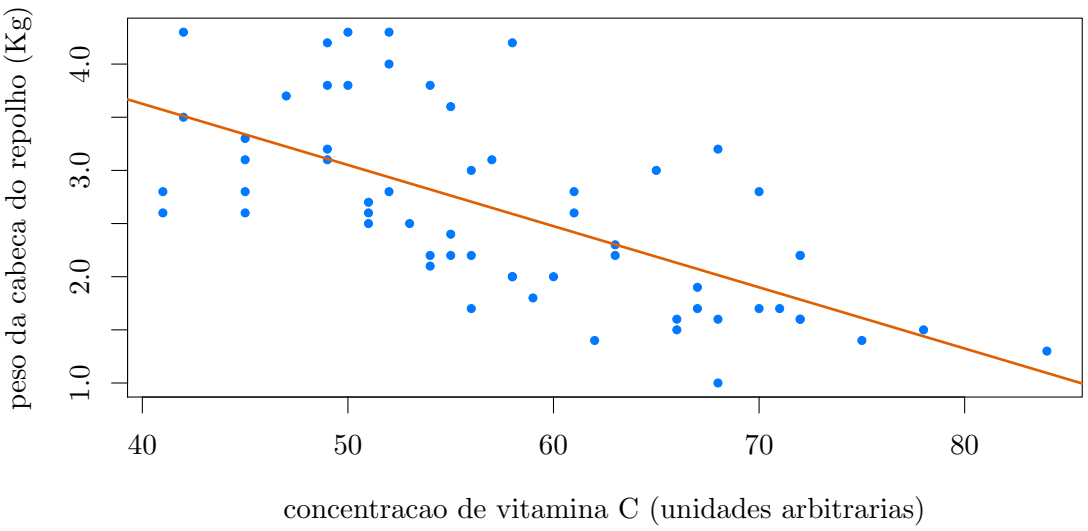

Figura 4: regressão linear para os dados do Exercício 80.

Exercício 81. Para verificar se existe relação entre a renda familiar (em salários mínimos) e o número de filhos, foi coletada uma amostra de 8 famílias em uma ciudade. Os resultados obtidos são apresentados na seguinte tabela.

Família	1	2	3	4	5	6	7	8
Renda	12	14	15	17	23	27	34	43
Filhos	3	2	2	1	1	0	0	0

(i) Que conclusões podem ser tiradas, baseando-se em um diagrama de dispersão, apresentado acima, e no coeficiente de correlação? (ii) Calcule a reta de mínimos quadrados e interprete os parâmetros. (iii) Verifique se a renda influi no número de filhos, utilizando $\alpha = 5\%$.

Exercício 82. Procure e carrege do site do curso os dados `Cars93.txt`. Utilize a função `read.table`. Estes dados contém 93 linhas e 27 colunas, e apresentam diversas características de vários automóveis americanos em 1993. Os dados foram tomados do pacote `MASS`, e podem ser carregados na memória aos escrever `library(MASS)`[6], caso este pacote esteja instalado na sua distribuição de R. Uma vez carregados os dados, digite `help(Cars93)` e também diretamente `Cars93` para obter maiores informações. O boxplot mostrado na figura 5 foi realizado com o comando `attach(Cars93);` e logo `boxplot(Price~Type,notch=F)`. (i) Baseado neste gráfico, você acredita que existe evidência para pensar que os preços médios dos vehículos variam de acordo ao tipo? (ii) O teste ANOVA para os preços dos veículos de

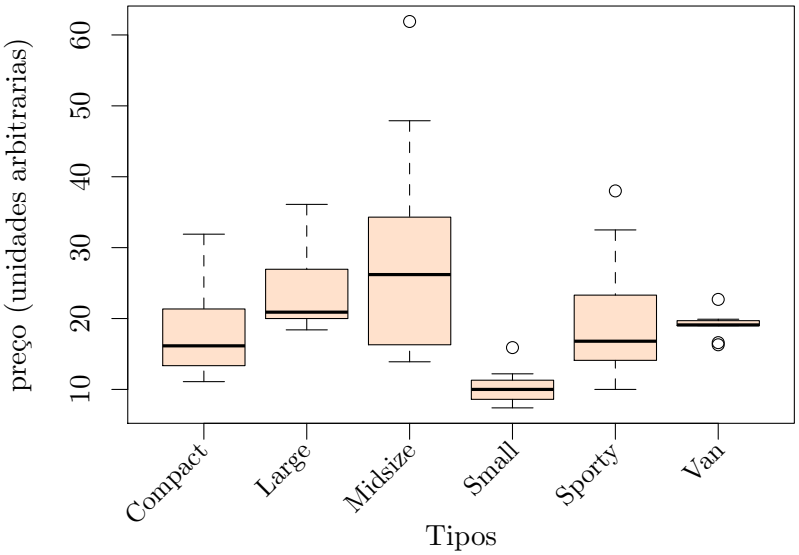

Figura 5: preços de diversos tipos de carros americanos em 1993.

acordo as classes em `Types` pode ser realizado como

```
anova(lm(Price~Type))
```

resultando

```
Analysis of Variance Table

Response: Price
          Df Sum Sq Mean Sq F value    Pr(>F)
Type       5 3421.4   684.3  11.532 1.477e-08 ***
Residuals 87 5162.6    59.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Em base a este teste podemos descartar a hipótese que consiste em pensar que todos os tipos de carros apresentam o mesmo valor médio? (iii) Faça uma regressão linear utilizando `Weight` como variável independente e `MPG.highway`. Qual é o resultado do teste F associado? (iv) Considere o teste

```
t.test(Price~Origin, alternative=two.sided)
```

[6]MASS contém os dados e as funções que acompanham a referéncia: Venables, W. N. e Ripley, B. D. (1999) *Modern Applied Statistics with S-PLUS*. Terceira Edição. Springer Verlag.

onde `Origin` e uma variável com dois valores `USA` e `non-USA`. O que esta sendo testado (quais são H_0 e H_a)? Qual é o resultado do teste? (v) Considere o teste

```
t.test(Price~Origin, alternative=greather)
```

Quais são as hipoteses? Qual é o resultado do teste? (veja como muda a conclusão do teste em `alternative hypotesis`).

Exercício 83. Um hospital deseja verificar o grau de satisfação de seus pacientes. Para tanto, escolheu domicílios de famílias de classe X, Y e Z, que fizeram uso do hospital, e solicitou que um questionário fosse preenchido. Os questionários foram devidamente codificados, a fim de fornecer um índice de satisfação que varia de 1 a 5 (insatisfeito a satisfeito). Os resultados do questionario se encontram no aquivo `hospital.txt`. Faça um teste ANOVA para verificar se o índice de satisfação médio varia ou não de classe a classe. Qual é a conclusão se α =0.05%?

5 Dados categóricos

Exercício 84. Um dado é jogado 180 vezes com os seguintes resultados

valor da face superior	1	2	3	4	5	6
frequência	28	36	36	30	28	23

Esse dado é balanceado? Use um nível de significância de 0.01.

Exercício 85. Um teste destinado a medir o nível de ansiedade foi aplicado a uma amostra homens (h) e outra formada por mulheres (m) antes de sofrer o mesmo procedimento cirúrgico. Os tamanhos das amostras e os desvios calculados a partir dos resultados são os seguintes,

$$n_h = 16,\ s_h^2 = 150; \qquad n_m = 21,\ s_m^2 = 275.$$

Diga se esses dados fornecem evidências suficientes para indicar que a amostra das mulheres é mais variável do que a amostra dos homens. Considere $\alpha = 0.05$.

Exercício 86. Uma amostra aleatória de 200 homens casados, todos aposentados, foi classificada de acordo com o nível educacional e o número de filhos.

	Número de filhos		
Educação	0-1	2-3	Acima de 3
Elementar	14	37	32
Média	19	42	17
Superior	12	17	10

Teste a hipótese de que, num nível de significância de 0.05, o tamanho da família é independente do nível de educação obtido pelo pai.

Exercício 87. Em um experimento para verificar a relação entre crises de asma e incidência de gripe, 150 crianças foram escolhidas, ao acaso, dentre aquelas acompanhadas pelo HC. Os dados referentes a uma semana são apresentados na seguinte tabela.

Asma\Gripe	Sim	Não
Sim	27	34
Não	42	47

Teste se a ocorrência de asma e gripe são independentes.

Exercício 88. Suponha que devemos determinar se as opiniões de eleitores residentes no estado de São Paulo sobre a reforma tributária são independentes de seus níveis de renda. Mil eleitores em uma amostra aleatória foram classificados como eleitores de baixa, média ou alta renda e se a favor ou não da reforma. As frequências observadas são apresentadas na tabela abaixo. Qual é a clonclusão ao nível de 5%?

	Renda		
	Baixa	Média	Alta
favorável	182	213	203
contrário	154	138	110

Exercício 89. A temporada de gripe no município de São Joaquim (Santa Catarina) ocorre entre maio e setembro, os meses mais frios do ano. A Secretaria Municipal de Saúde informou os seguintes dados para as ocorrências em 2014.

mês	número total de casos de gripe
maio	62
junho	84
julho	17
agosto	16
setembro	21
Total	200

Temos interesse em saber se o número de casos de gripe são distribuídos igualmente entre os cinco meses mais frios do ano. Ou seja, queremos saber se os casos de gripe seguem uma distribuição uniforme ao longo do tempo. Qual é a conclussão ao nível $\alpha = 0.001$?

Exercício 90. Uma moeda é lançada até obtermos a primeira cara. Seja X o número de lançamentos necessários. Os resultados obtidos após de repetir o experimento 256 vezes é fornecido pela seguinte tabela,

x	1	2	3	4	5	6	7	8
frequência	136	60	34	12	9	1	3	1

Teste ao nível de 0,05 se a distribuição de X pode ser ajustada pela distribuição Geometrica(p) com $p = \frac{1}{2}$. X é Geometrica(p) se $P(X = x) = (1-p)^{x-1}p$, $x = 1, 2, \ldots$. Nesse caso $\mathbb{E}[X] = \frac{1}{p}$.

6 Apêndice

6.1 Distribuições amostrais

Esta seção apresenta diversos resultados sobre a origem de varias distribuições amostrais utilizadas em aula. O seu estudo é opcional.

6.1.1 Distribuições Gamma e χ^2

Apresentamos duas distribuições essenciais no estudo da distribuição amostral de S^2.

Definição 1. A variável aleatória X tem distribuição gamma com parâmetros α e $\beta > 0$ se a sua densidade é dada por

$$f_X(x) = \begin{cases} \dfrac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & \text{se } x \geq 0, \\ 0, & \text{caso contrário} \end{cases}$$

sendo $\Gamma : (0, +\infty) \to [0, +\infty)$ a função definida pela integral $\Gamma(x) = \displaystyle\int_0^{+\infty} t^{x-1} e^{-t} dt$.

Lema 1. *Se X é normal padrão, então X^2 tem distribuição gamma com parâmetros $\alpha = 1/2$ e $\beta = 2$.*

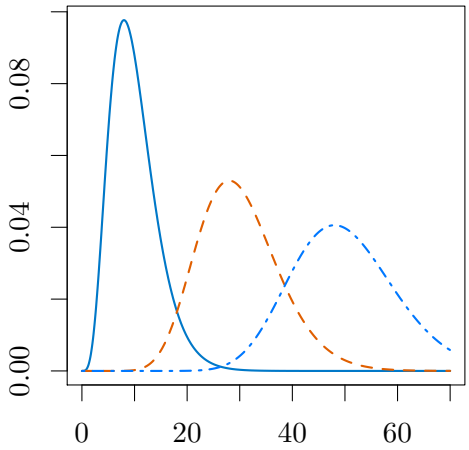

Figura 6: densidade χ^2 para 10 (linha continua), 30 e 50 graus de liberdade.

Lema 2. *Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes gamma com parâmetros α_i, β respectivamente. A variável aleatória $X_1 + \ldots + X_n$ tem distribuição gamma com parâmetros $\alpha_1 + \ldots + \alpha_n$ e β.*

Suponhamos agora que $X_1, \ldots, X_n$ é uma amostra *i.i.d.* de uma população normal padrão. Neste caso diante ao exposto temos que $X_1^2, \ldots, X_n^2$ são independentes e com distribuição gamma com $\alpha = 1/2$ e $\beta = 2$. Do Lema 2 temos que

$$X_1^2 + \ldots + X_n^2 \sim \text{gamma}\Big(\frac{n}{2}, 2\Big). \tag{7}$$

Definição 2. Uma variável aleatória tem distribuição χ^2, "Qi-quadrado", com n graus de liberdade se esta tem distribuição gamma com parâmetros $\alpha = n/2$ e $\beta = 2$.

Esta terminologia introduzida pelo estatístico Britânico K. Pearson (1857-1936) ainda é utilizada hoje em dia. A figura 6 mostra a densidade χ^2 para diferentes graus de liberdade.

Lema 3. *(i) Sejam X e Y independentes e distribuidas de acordo a distribuição χ^2 com n e m graus de liberdade respectivamente. $X + Y$ tem distribuição χ^2 com $n + m$ graus de liberdade. (ii) Se X e $X + Y$ são χ^2 com m e n, $m < n$, graus de liberdade, então Y é χ^2 com $n - m$ graus de liberdade.*

O seguinte Teorema permite obter a distribuição amostral do estatístico S^2, rescalado pela constante $(n-1)/\sigma^2$, quando são consideradas amostras *i.i.d.* de uma população normal com variância σ^2.

Teorema 1. *Seja $X_1, \ldots, X_n$, $n \geq 2$, uma amostra i.i.d. de uma população normal com média μ e variância σ^2. A variável aleatória*

$$V = \frac{(n-1)S^2}{\sigma^2}$$

apresenta distribuição χ^2 com $n - 1$ graus de liberdade.

6.1.2 Distribuição t (t-Student)

Apresentamos a distribuição da variável aleatória

$$T = \sqrt{n}\Big(\frac{\bar{X} - \mu}{S}\Big)$$

obtida ao considerar uma amostra *i.i.d.* de uma população normal.

Teorema 2. *Seja Z normalmente distribuída com média e variância 0 e 1 respectivamente, e seja V com distribuição χ^2 com n graus de liberdade. Se Z e V são independentes, então a variável aleatória*

$$T = \frac{Z}{\sqrt{V/n}}$$

tem densidade de probabilidade f dada por

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})}\Big(1 + \frac{x^2}{n}\Big)^{-\frac{n+1}{2}} \qquad \textit{para todo } x \in \mathbb{R}. \tag{8}$$

Definição 3. Uma variável aleatória tem distribuição t com n graus de liberdade se a sua densidade é dada pela lei em (8).

A distribuição t foi descrita inicialmente por William S. Gosset (1876-1937). Gosset trabalhava na cervejaria Guiness em Dublim a qual proibia que os seus empleados publicassem o seu trabalho científico. Devido a isto Gosset publico os seus trabalhos utilizando o pseudónimo "Student". Em honra ao seu descobridor hoje em dia a distribuição t também é conhecida como a "distribuição Student" (ou t-Student). Gráficos da densidade t são apresentados na figura 7.

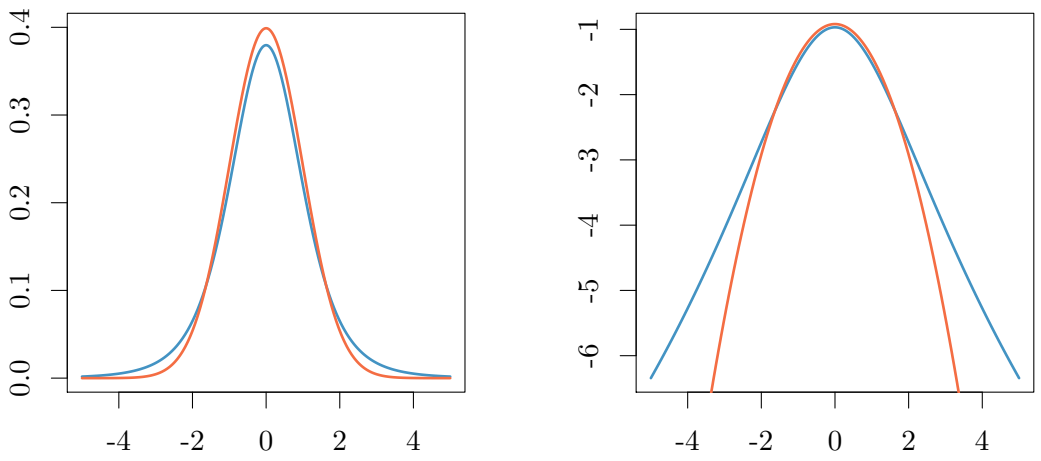

Figura 7: Esquerda: densidade t de Student para 5 graus e densidade normal padrão. A figura a direita apresenta o logarirmo da densidade, com o objetivo de evidenciar as diferenças entre ambos modelos. A densidade t apresenta caudas 'mais pesadas' do que a densidade normal.

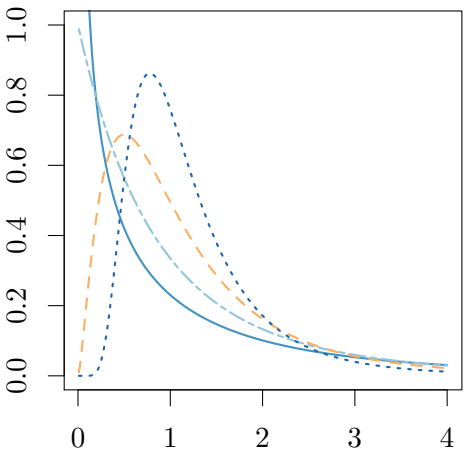

Figura 8: densidades $F(m, n)$ para vários valores de m e n (linha continua $(1, 10)$, $(2, 10)$, $(5, 10)$, e $(30, 10)$).

6.1.3 Distribuição F

Sejam X e Y duas populações e S_X^2, S_Y^2 os estimadores das variâncias σ_X^2 e σ_Y^2. Desejamos estudar o quociente σ_X^2/σ_Y^2 e a tal fim determinamos a distribuição de

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}.$$

Esta variável aleatória tem "distribuição F".

Definição 4. A variável aleatória X apresenta distribuição F com m graus de liberdade no numerados e n graus de liberdade no denominador se a sua densidade é dada por

$$f(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{m}{n}\right)^{m/2} x^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}}, & \text{se } x > 0, \\ 0, & \text{se } x \leq 0. \end{cases}$$

A distribuição F é também conhecida como a distribuição de Fisher em honra a Sir Ronald A. Fisher (1890–1962), quem mostrou o seguinte resultado chave para determinarmos a distribuição amostral de σ_X/σ_Y.

Teorema 3. *Sejam U e V duas variáveis aleatórias com distribuição χ^2 de m e n graus de liberdade respectivamente. Se U e V são independentes, então*

$$\frac{U/m}{V/n}$$

tem distribuição F com m graus de liberdade no numerador e n graus de liberdade no denominador.

6.2 Convergência de variáveis aleatórias

Apresentamos vários resultados relativos a convergência de variáveis aleatórias utilizados durante o curso. As pessoas interessadas podem encontrar as demosntrações da maior parte destes resultados em [3] ou [2].

Definição 5. Sejam (X_n), $n \geq 1$, e X, variáveis aleatórias definidas no mesmo espaço de probabilidade $(\Omega, \mathfrak{B}, \mathbb{P})$ e sejam F_{X_n} e F_X as suas funções de distribuição.
(i) X_n converge a X em probabilidade, denotado $X_n \xrightarrow{\mathbb{P}} X$, se para todo $\varepsilon > 0$,

$$\mathbb{P}\big(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\big) \to 0, \quad \text{quando } n \to \infty.$$

(ii) X_n converge em distribuição a X, denotado $X_n \xrightarrow{\mathfrak{D}} X$, se

$$F_n(x) \to F(x) \quad \text{quando } n \to \infty, \text{ para todo } x \in \mathbb{R} \text{ onde } F(x) \text{ é continua.}$$

(iii) X_n converge quase certamente a X, denotado $X_n \xrightarrow{q.c.} X$, se

$$\mathbb{P}\Big(\Big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\}\Big) = 1.$$

Exemplos clásicos da convergência em (i) e (iii) são fornecidos pela Lei Fraca e a Lei Forte dos Grandos Numeros, respectivamente. Um exemplo de (ii) é constituído pelo Teorema Central do Limite.

6.3 Leis dos Grandes Números

Seja X_n, $n \in \mathbb{N}$, uma sequência de variáveis aleatórias independentes, e seja $S_n = \sum_{i=0}^{n} X_i$ a sua soma parcial. Em esta seção estudaremos o comportamento de S_n no limite quando $n \to \infty$. Em geral, é possível formular o problema da seguinte maneira. Se a_n e b_n são duas sequências de númeors reais, quais são as condições que garantem o limite

$$S_n/b_n - a_n \longrightarrow S \quad \text{quando} \quad n \to \infty \tag{9}$$

sendo "$\longrightarrow$" uma das formas de convergência mencionadas na definição 1. Esta seção descreve um resultado fundamentais conhecido como a Lei Fraca dos Grandes Números, no qual a convergência é em probabilidade.

6.3.1 Lei Fraca dos Grandes Números

Lema 4 (Desigualdade de Chebyshev)**.** *Se X é uma variável aleatória integrável, então para qualquer constante $k > 0$*

$$\mathbb{P}\big(|X - \mathbb{E}[X]| \geq k\big) \leq \frac{\mathrm{Var}(X)}{k^2}$$

Demonstração. Veja [2]. □

Teorema 4 (Lei Fraca dos Grandes Números. Chebyshev, 1867)**.** *Seja $X_1, X_2, \ldots$ uma sequência de variáveis aleatórias independentes, e seja S_n a sua soma parcial até n. Se para todo n, $\mathrm{Var}(X_n) \leq K$ onde K é uma constante finita, então*

$$\frac{S_n - \mathbb{E}[S_n]}{n} \xrightarrow{\mathbb{P}} 0.$$

Demonstração. Devemos mostrar que para qualquer $\varepsilon > 0$, $\mathbb{P}(|S_n - \mathbb{E}[S_n]|/n \geq \varepsilon) \to 0$ quando $n \to \infty$. Pelas hipóteses do enunciado temos $\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) \leq nK$, logo da desigualdade (clássica) de Chebyshev

$$\mathbb{P}\big(|S_n - \mathbb{E}[S_n]| \geq \varepsilon n\big) \leq \frac{\mathrm{Var}(S_n)}{\varepsilon^2 n^2} \leq \frac{K}{\varepsilon^2 n} \to 0.$$ □

Exemplo 1 (Ensaios Bernoulli)**.** Apresentamos um exemplo simples porem importante para desenvolver a nossa intuição. O seguinte exemplo é de fato a primeira Lei dos Grandes Números publicada em 1713, após de 8 anos da morte de J. Bernoulli, [1]. Suponhamos que lançamos uma moeda n vezes, e neste caso consideramos a sequencia de variáveis aleatórias $\xi_1, \ldots, \xi_n$, tais que para $1 \leq i \leq n$, $\xi_i(\omega) = \mathbf{1}_{\text{Cara}}(\omega_i)$, ou seja, $\xi_i = 1$ se o i-ésimo lançamento resulta em cara, e $\xi_i = 0$ no caso contrário (se o resultado é coroa). Assim $S_n = \sum_{i=1}^{n} \xi_i$, o número de caras em n lançamentos, é uma variável aleatória Binomial(n, p), onde $p = \mathbb{P}(\xi_i = 1)$ é a probabilidade de sair cara em qualquer lançamento. Temos portanto que $\mathbb{E}[S_n] = np$, logo $\mathbb{E}[S_n/n] = p = \mathbb{E}[\xi_i]$. A ley dos grandes números neste caso afirma que

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} p. \tag{10}$$

Este resultado é conhecido como a Ley dos Grandes Números para ensaios Bernoulli.

A figura figura 9 apresenta várias instâncias do experimento do lançamento da moeda com o objetivo de visualizar a convergência em (10). A linha contínua apresenta um dos possiveis resultados ao lançar $n = 250$ vezes uma moeda viciada com $p = 0.2$. Os valores para ξ_i em cada lançamento sao apresentados por circulos, e S_i/i, $1 \leq i \leq n$, pela linha continua. Os valores de S_i/i são apresentados para quatro outras possíveis realizações ω do experimento. Claramente, a figura mostra que, independentemente da realização, S_n/n se aproxima do valor de p a medida que n aumenta.

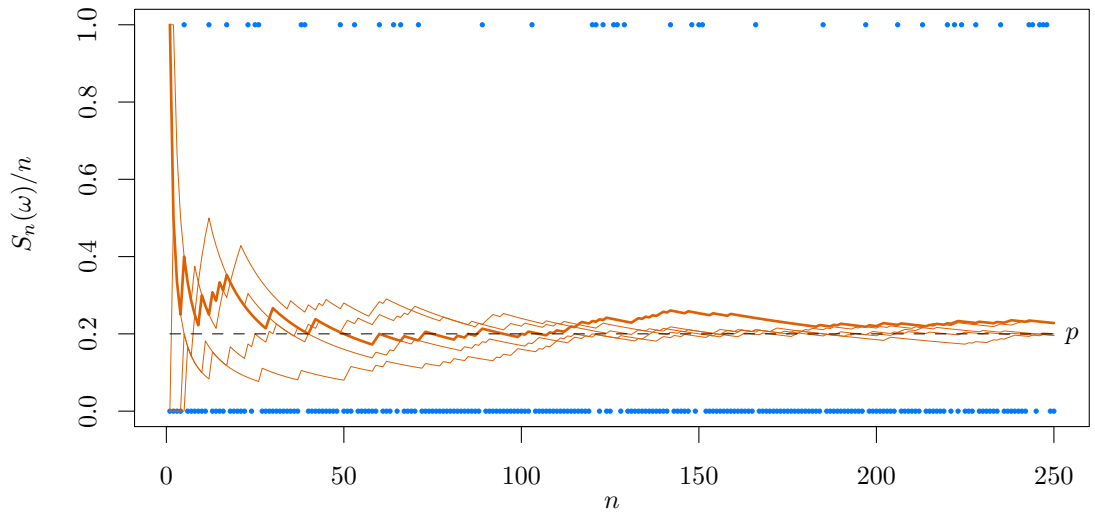

Figura 9: varias simulações de 250 lançamentos de uma moeda viciada com $\mathbb{P}(\{\text{cara}\}) = p = 0.2$. A sequência de caras e coroas para a primeira simulação, ω^1, corresponde aos circulos em 0 (coroa) e em 1 (cara). A linha continua representa os valores de $S_n(\omega^1)/n$, e as otras linhas correspondem aos valores para três outras realizações do processo, ω^2, ω^3, ω^4.

É possível obter uma Lei Fraca sem assumir que as variâncias das variáveis X_n sejam finitas. Embora, esta hipótese é crucial para a Lei Fraca de Chebyshev apresentada no Teorema 4.

Teorema 5 (Lei Fraca dos Grandes Numeros. Khintchin, 1929)**.** *Sejam X_1, X_2, ... variáveis aleatórias independentes e identicamente distribuídas com média finita μ. Se S_n denota a soma parcial de X_n, então*

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu.$$

Demonstração. Veja [5]. □

6.4 Teorema Central do Limite

Passamos agora a estudar a convergência da distribuição de S_n, quando S_n é corretamente rescalada. Em geral veremos como sob certas hipoteses é possível estabelecer que

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}(S_n)}} \leq x\right) = \Phi(x), \quad x \in \mathbb{R},$$

onde

$$\Phi(x) = \int_{-\infty}^{x} \phi(u)du, \qquad \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{11}$$

isto é, ϕ denota a densidade de probabilidade normal (com média 0 e variância 1).

Teorema 6 (Teorema Central do Limite de Lindenberg-Lévy)**.** *Sejam X_1, X_2, ... variáveis aleatórias independentes e identicamente distribuídas, tais que $\mathbb{E}[X_1] = \mu$, e $Var(X_1) = \sigma^2 < \infty$. Seja $S_n = \sum_{i=1}^n X_i$, e Z uma variável aleatória normal com média 0 e variância 1, então*

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathfrak{D}} Z.$$

Demonstração. A prova deste Teorema pode ser encontrada em [2], p. 194 ou em [3], p. 240. □

O seguinte resultado mostra que o Teorema Central do Limite é válido ainda quando as variáveis aleatórias X_1, X_2, ..., não apresentam a mesma distribuição.

Teorema 7 (Theorema Central do Limite. Kolmogorov, 1933)**.** *Seja X_1, X_2, ... uma sequência de variáveis aleatórias independentes, e seja S_n a sua soma parcial. Para cada i sejam $\mu_i = \mathbb{E}[X_i]$, e $\sigma_i^2 = \mathrm{Var}(X_i)$, logo $m_n = \sum_{i=1}^n \mu_i$ e $s_n^2 = \sum_{i=1}^n \sigma_i^2$ denotam a média e a variância de S_n, e seja X uma variável aleatória normal com média 0 e variância 1. Sob as seguentes hipóteses adicionais*

(i) $s_n^2 \to \infty$ quando $n \to \infty$,

(ii) existe uma constante K, tal que para todo i, $\mathbb{P}(|X_i| \leq K) = 1$,

tem-se

$$\frac{S_n - m_n}{s_n} \xrightarrow{\mathfrak{D}} X.$$

Referências

[1] J. Bernoulli. *...Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis.* Impensis Thurnisiorum, fratrum, Basileae, 1713.
Tradução: E. D. Sylla. *The Art of Conjecturing, together with Letter to a friend of Sets in Court Tennis.* The Johns Hopkins University Press, 2005.

[2] G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Oxford University Press, Oxford, UK, 3rd edition, 2001.

[3] B. R. James. *Probabilidade: um curso em nível intermediário.* Projeto Euclides. Associação Instituto Nacional de Matemática Pura e Aplicada, Rio de Janeiro, 2002.

[4] M. Nascimento Magalhães and A. C. Pedroso de Lima. *Noções de Probabilidade e Estatística.* Edusp, São Paulo, 6a edition, 2004.

[5] C. R. Rao. *Linear Statistical Inference and its Applications*. Wiley, New York, 1973.

[6] T. R. T. R. Golub, D. K. Slonim, P. Amayo, D. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, M. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

7 Tabelas

Tabela 2: valores da distribuição normal padrão. A tabela frnece os valores de z definidos por α tais que $\alpha = \mathbb{P}(0 \leq Z \leq z)$. Colunas correspondem a segunda casa decimal de z e linhas a parte inteira e primeira casa decimal.

	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Tabela 3: Valores da distribuição t-Student bicaudal. A tabela fornece os valores de x para α, onde $\alpha = \mathbb{P}(|T| \geq x)$, ou alternativamente para γ onde $\gamma = 1 - \alpha = \mathbb{P}(-x < T < x)$. GL denota os graus de liberdade.

GL	γ	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.98	0.99	0.995	0.998	0.999
	α	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
1		0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.82	63.657	127.321	318.31	636.6
2		0.617	0.817	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.6
3		0.584	0.765	0.979	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.92
4		0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5		0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6		0.553	0.718	0.910	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7		0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8		0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.897	3.355	3.833	4.501	5.041
9		0.544	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10		0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11		0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12		0.539	0.696	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13		0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14		0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.625	2.977	3.326	3.787	4.140
15		0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16		0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.584	2.921	3.252	3.686	4.015
17		0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18		0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19		0.533	0.688	0.861	1.066	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21		0.533	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22		0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23		0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24		0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.090	3.467	3.745
25		0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26		0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27		0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28		0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29		0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30		0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31		0.530	0.683	0.853	1.054	1.309	1.695	2.040	2.453	2.744	3.022	3.375	3.633
32		0.530	0.682	0.853	1.054	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33		0.530	0.682	0.853	1.053	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34		0.529	0.682	0.852	1.052	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35		0.529	0.682	0.852	1.052	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36		0.529	0.681	0.852	1.052	1.306	1.688	2.028	2.434	2.719	2.991	3.333	3.582
37		0.529	0.681	0.851	1.051	1.305	1.687	2.026	2.431	2.715	2.985	3.326	3.574
38		0.529	0.681	0.851	1.051	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566
39		0.529	0.681	0.851	1.050	1.304	1.685	2.023	2.426	2.708	2.976	3.313	3.558
40		0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
42		0.528	0.680	0.850	1.049	1.302	1.682	2.018	2.418	2.698	2.963	3.296	3.538
44		0.528	0.680	0.850	1.049	1.301	1.680	2.015	2.414	2.692	2.956	3.286	3.526
46		0.528	0.680	0.850	1.048	1.300	1.679	2.013	2.410	2.687	2.949	3.277	3.515
48		0.528	0.680	0.849	1.048	1.299	1.677	2.011	2.407	2.682	2.943	3.269	3.505
50		0.528	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60		0.527	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
70		0.527	0.678	0.847	1.044	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
80		0.527	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
90		0.526	0.677	0.846	1.042	1.291	1.662	1.987	2.369	2.632	2.878	3.183	3.402
100		0.526	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.391
120		0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
150		0.526	0.676	0.844	1.040	1.287	1.655	1.976	2.351	2.609	2.849	3.145	3.357
200		0.525	0.676	0.843	1.039	1.286	1.652	1.972	2.345	2.601	2.839	3.131	3.340
300		0.525	0.675	0.843	1.038	1.284	1.650	1.968	2.339	2.592	2.828	3.118	3.323
500		0.525	0.675	0.842	1.038	1.283	1.648	1.965	2.334	2.586	2.820	3.107	3.310
∞		0.524	0.675	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Tabela 4: Distribuição χ^2. A tabela fornece o valor de x para α de maneira que $\mathbb{P}(\chi^2 \geq x) = \alpha$. GL denota graus de liberdade.

α / GL	.99	.98	.975	.95	.9	.8	.7	.5	.3	.2	.05	.04	.025	.02	.01	.002	.001
1	0	0.001	0.001	0.004	0.016	0.064	0.148	0.455	1.074	1.642	3.841	4.218	5.024	5.412	6.635	9.55	10.828
2	0.02	0.04	0.051	0.103	0.211	0.446	0.713	1.386	2.408	3.219	5.991	6.438	7.378	7.824	9.21	12.429	13.816
3	0.115	0.185	0.216	0.352	0.584	1.005	1.424	2.366	3.665	4.642	7.815	8.311	9.348	9.837	11.345	14.796	16.266
4	0.297	0.429	0.484	0.711	1.064	1.649	2.195	3.357	4.878	5.989	9.488	10.026	11.143	11.668	13.277	16.924	18.467
5	0.554	0.752	0.831	1.145	1.61	2.343	3	4.351	6.064	7.289	11.07	11.644	12.833	13.388	15.086	18.907	20.515
6	0.872	1.134	1.237	1.635	2.204	3.07	3.828	5.348	7.231	8.558	12.592	13.198	14.449	15.033	16.812	20.791	22.458
7	1.239	1.564	1.69	2.167	2.833	3.822	4.671	6.346	8.383	9.803	14.067	14.703	16.013	16.622	18.475	22.601	24.322
8	1.646	2.032	2.18	2.733	3.49	4.594	5.527	7.344	9.524	11.03	15.507	16.171	17.535	18.168	20.09	24.352	26.124
9	2.088	2.532	2.7	3.325	4.168	5.38	6.393	8.343	10.656	12.242	16.919	17.608	19.023	19.679	21.666	26.056	27.877
10	2.558	3.059	3.247	3.94	4.865	6.179	7.267	9.342	11.781	13.442	18.307	19.021	20.483	21.161	23.209	27.722	29.588
11	3.053	3.609	3.816	4.575	5.578	6.989	8.148	10.341	12.899	14.631	19.675	20.412	21.92	22.618	24.725	29.354	31.264
12	3.571	4.178	4.404	5.226	6.304	7.807	9.034	11.34	14.011	15.812	21.026	21.785	23.337	24.054	26.217	30.957	32.909
13	4.107	4.765	5.009	5.892	7.042	8.634	9.926	12.34	15.119	16.985	22.362	23.142	24.736	25.472	27.688	32.535	34.528
14	4.66	5.368	5.629	6.571	7.79	9.467	10.821	13.339	16.222	18.151	23.685	24.485	26.119	26.873	29.141	34.091	36.123
15	5.229	5.985	6.262	7.261	8.547	10.307	11.721	14.339	17.322	19.311	24.996	25.816	27.488	28.259	30.578	35.628	37.697
16	5.812	6.614	6.908	7.962	9.312	11.152	12.624	15.338	18.418	20.465	26.296	27.136	28.845	29.633	32	37.146	39.252
17	6.408	7.255	7.564	8.672	10.085	12.002	13.531	16.338	19.511	21.615	27.587	28.445	30.191	30.995	33.409	38.648	40.79
18	7.015	7.906	8.231	9.39	10.865	12.857	14.44	17.338	20.601	22.76	28.869	29.745	31.526	32.346	34.805	40.136	42.312
19	7.633	8.567	8.907	10.117	11.651	13.716	15.352	18.338	21.689	23.9	30.144	31.037	32.852	33.687	36.191	41.61	43.82
20	8.26	9.237	9.591	10.851	12.443	14.578	16.266	19.337	22.775	25.038	31.41	32.321	34.17	35.02	37.566	43.072	45.315
21	8.897	9.915	10.283	11.591	13.24	15.445	17.182	20.337	23.858	26.171	32.671	33.597	35.479	36.343	38.932	44.522	46.797
22	9.542	10.6	10.982	12.338	14.041	16.314	18.101	21.337	24.939	27.301	33.924	34.867	36.781	37.659	40.289	45.962	48.268
23	10.196	11.293	11.689	13.091	14.848	17.187	19.021	22.337	26.018	28.429	35.172	36.131	38.076	38.968	41.638	47.391	49.728
24	10.856	11.992	12.401	13.848	15.659	18.062	19.943	23.337	27.096	29.553	36.415	37.389	39.364	40.27	42.98	48.812	51.179
25	11.524	12.697	13.12	14.611	16.473	18.94	20.867	24.337	28.172	30.675	37.652	38.642	40.646	41.566	44.314	50.223	52.62
26	12.198	13.409	13.844	15.379	17.292	19.82	21.792	25.336	29.246	31.795	38.885	39.889	41.923	42.856	45.642	51.627	54.052
27	12.879	14.125	14.573	16.151	18.114	20.703	22.719	26.336	30.319	32.912	40.113	41.132	43.195	44.14	46.963	53.023	55.476
28	13.565	14.847	15.308	16.928	18.939	21.588	23.647	27.336	31.391	34.027	41.337	42.37	44.461	45.419	48.278	54.411	56.892
29	14.256	15.574	16.047	17.708	19.768	22.475	24.577	28.336	32.461	35.139	42.557	43.604	45.722	46.693	49.588	55.792	58.301
30	14.953	16.306	16.791	18.493	20.599	23.364	25.508	29.336	33.53	36.25	43.773	44.834	46.979	47.962	50.892	57.167	59.703

Tabela 5: Distribuição F-Fisher(n, d). A tabela fornece o valor de $x = F_\alpha(n, d)$, ou seja, x tal que $\mathbb{P}(F \geq x) = 0.05$ quando F é uma variável aleatória Fisher com n graus de liberdade no numerador e d no denominador. Colunas correspondem aos graus de liberdade n e linhas aos graus de liberdade d.

n	1	2	3	4	5	6	7	8	9	10	11	12
1	161.448	199.5	215.707	224.583	230.162	234	236.768	238.883	240.543	241.882	242.983	243.9
2	18.513	19	19.164	19.247	19.296	19.33	19.353	19.371	19.385	19.396	19.5	19.41
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.763	8.745
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.936	5.912
5	6.608	5.786	5.409	5.192	5.05	4.95	4.876	4.818	4.772	4.735	4.704	4.678
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.06	4.027	4
7	5.591	4.737	4.347	4.12	3.972	3.866	3.787	3.726	3.677	3.637	3.603	3.575
8	5.318	4.459	4.066	3.838	3.687	3.581	3.5	3.438	3.388	3.347	3.313	3.284
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.23	3.179	3.137	3.102	3.073
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.02	2.978	2.943	2.913
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.818	2.788
12	4.747	3.885	3.49	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.717	2.687
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.635	2.604
14	4.6	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.565	2.534
15	4.543	3.682	3.287	3.056	2.901	2.79	2.707	2.641	2.588	2.544	2.507	2.475
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.456	2.425
17	4.451	3.592	3.197	2.965	2.81	2.699	2.614	2.548	2.494	2.45	2.413	2.381
18	4.414	3.555	3.16	2.928	2.773	2.661	2.577	2.51	2.456	2.412	2.374	2.342
19	4.381	3.522	3.127	2.895	2.74	2.628	2.544	2.477	2.423	2.378	2.34	2.308
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.31	2.278
25	4.242	3.385	2.991	2.759	2.603	2.49	2.405	2.337	2.282	2.236	2.198	2.165
30	4.171	3.316	2.922	2.69	2.534	2.421	2.334	2.266	2.211	2.165	2.126	2.092
60	4.001	3.15	2.758	2.525	2.368	2.254	2.167	2.097	2.04	1.993	1.952	1.917
120	3.92	3.072	2.68	2.447	2.29	2.175	2.087	2.016	1.959	1.91	1.869	1.834

n	13	14	15	16	17	18	19	20	25	30	60	120
1	244.69	245.364	245.95	246.464	246.918	247.32	247.686	248.013	249.26	250.095	252.196	253.3
2	19.419	19.424	19.429	19.433	19.437	19.44	19.443	19.446	19.456	19.462	19.5	19.5
3	8.729	8.715	8.703	8.692	8.683	8.675	8.667	8.66	8.634	8.617	8.572	8.549
4	5.891	5.873	5.858	5.844	5.832	5.821	5.811	5.803	5.769	5.746	5.688	5.658
5	4.655	4.636	4.619	4.604	4.59	4.579	4.568	4.558	4.521	4.496	4.431	4.398
6	3.976	3.956	3.938	3.922	3.908	3.896	3.884	3.874	3.835	3.808	3.74	3.705
7	3.55	3.529	3.511	3.494	3.48	3.467	3.455	3.445	3.404	3.376	3.304	3.267
8	3.259	3.237	3.218	3.202	3.187	3.173	3.161	3.15	3.108	3.079	3.005	2.967
9	3.048	3.025	3.006	2.989	2.974	2.96	2.948	2.936	2.893	2.864	2.787	2.748
10	2.887	2.865	2.845	2.828	2.812	2.798	2.785	2.774	2.73	2.7	2.621	2.58
11	2.761	2.739	2.719	2.701	2.685	2.671	2.658	2.646	2.601	2.57	2.49	2.448
12	2.66	2.637	2.617	2.599	2.583	2.568	2.555	2.544	2.498	2.466	2.384	2.341
13	2.577	2.554	2.533	2.515	2.499	2.484	2.471	2.459	2.412	2.38	2.297	2.252
14	2.507	2.484	2.463	2.445	2.428	2.413	2.4	2.388	2.341	2.308	2.223	2.178
15	2.448	2.424	2.403	2.385	2.368	2.353	2.34	2.328	2.28	2.247	2.16	2.114
16	2.397	2.373	2.352	2.333	2.317	2.302	2.288	2.276	2.227	2.194	2.106	2.059
17	2.353	2.329	2.308	2.289	2.272	2.257	2.243	2.23	2.181	2.148	2.058	2.011
18	2.314	2.29	2.269	2.25	2.233	2.217	2.203	2.191	2.141	2.107	2.017	1.968
19	2.28	2.256	2.234	2.215	2.198	2.182	2.168	2.155	2.106	2.071	1.98	1.93
20	2.25	2.225	2.203	2.184	2.167	2.151	2.137	2.124	2.074	2.039	1.946	1.896
25	2.136	2.111	2.089	2.069	2.051	2.035	2.021	2.007	1.955	1.919	1.822	1.768
30	2.063	2.037	2.015	1.995	1.976	1.96	1.945	1.932	1.878	1.841	1.74	1.683
60	1.887	1.86	1.836	1.815	1.796	1.778	1.763	1.748	1.69	1.649	1.534	1.467
120	1.803	1.775	1.75	1.728	1.709	1.69	1.674	1.659	1.598	1.554	1.429	1.352