UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Resolução de Exercícios de Estatística aplicados para as áreas de Ciências Biológicas e da Saúde

Programa de Monitoria da Graduação
Departamento de Estatística - UFMG
PMG-DEST

Ana Clara Gomes Rezende (Monitora)
Hanna Lury Umezu (Monitora)
Isabela Mendes Pinto Batista (Monitora)
Pedro Augusto Ribeiro Rios (Monitor)
Paulo Marcus Haratani Marques Meira (Monitor)
Vanessa de Melo Reis (Monitora)
Cristiano de Carvalho Santos (Orientador)

Belo Horizonte
2021

Sumário

1. Introdução

Essa apostila é composta por exercícios de estatística e de probabilidade aplicados à área das Ciências Biológicas e da Saúde. Todos os exercícios são acompanhados de sua resolução e interpretação dos resultados para que, assim, eles possibilitem uma melhor compreensão dos conteúdos abordados nas disciplinas de Bioestatística ofertadas pelo Departamento de Estatística da UFMG.

Nesse documento foi utilizado o software de R para resolução de muitos exercícios. O PMG-DEST já possui a apostila "Análises em Bioestatística Básica: uma introdução ao software R", disponível na página do programa dentro do site do Departamento de Estatística da UFMG, que apresenta mais detalhes sobre como instalar e utilizar o software.

O estudo de estatística é de suma importância no contexto das ciências, dado a constante produção de estudos. Para que estes estudos sejam desenvolvidos e interpretados de forma correta, métodos estatísticos adequados são essenciais ao longo de todas as suas etapas, como na coleta de dados, desenvolvimento de questionários e análise de resultados.

Essa apostila contém conteúdo de 3 módulos principais: Análise Descritiva de Dados, Probabilidade e Distribuição de Probabilidade.

É importante destacar que a maioria dos exercícios desta apostila foi realizada com **dados fictícios,** com o intuito de torná-los didáticos. Portanto, os resultados não devem ser extrapolados como dados reais.

2. Como utilizar os bancos de dados no programa R

Nessa apostila, para realização de alguns exercícios demonstrativos na área Ciências Biológicas e Saúde, foram utilizados alguns bancos de dados de pesquisas já existentes, bem como dados fictícios utilizados para fins didáticos. Todos esses bancos de dados estão disponíveis de forma gratuita, para acessá-los basta seguir o passo a passo:

Primeiramente é necessário instalar o pacote que contém os bancos de dados utilizados, chamado "aplore3", para isso utilize o comando install.packages:

"install.packages("aplore3")"

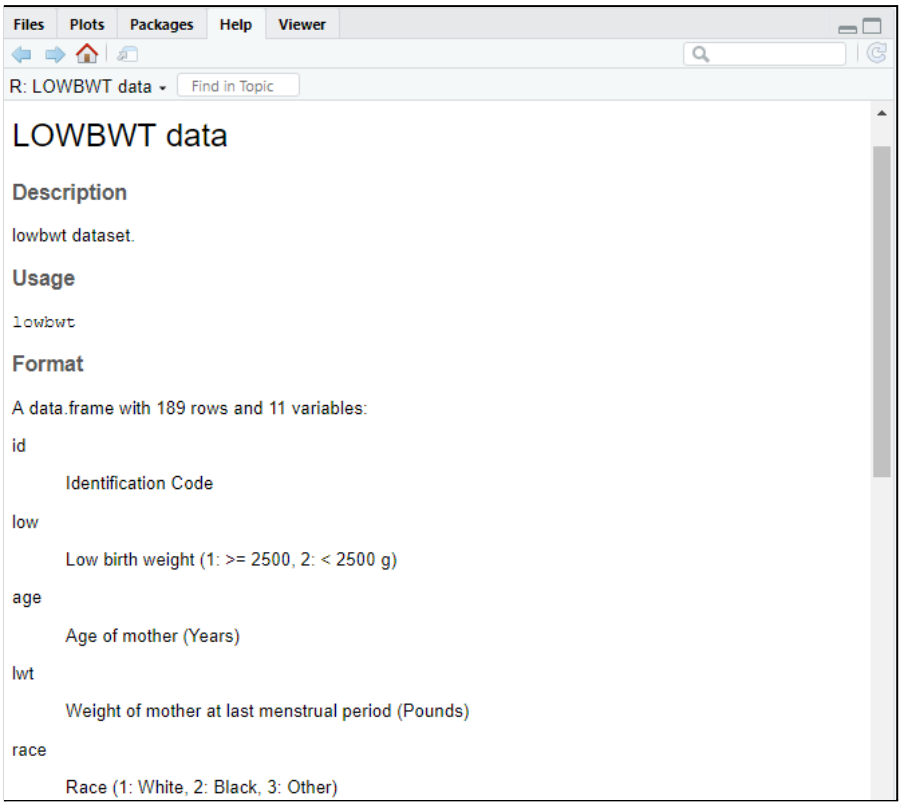
Assim que instalado, para poder utilizar os bancos de dados é necessário também carregar o pacote através do comando library:

"library(aplore3)"

Pronto, após essas etapas você já terá acesso aos bancos de dados utilizados nesta apostila, os quais são: “icu” - banco de dados sobre pacientes de UTI; “lowbwt” - banco de dados de um estudo sobre baixo peso em recém-nascidos; e “myopia” - banco de dados de um estudo sobre fatores de risco para o desenvolvimento de miopia em crianças.

Para se obter mais informações sobre os bancos de dados, como suas variáveis, basta utilizar o comando “?” seguido do nome do banco de dados de seu interesse, exemplo:

Assim, aparecerão em “help” as informações sobre esse banco de dados:



Na apostila de “Análises em Bioestatística Básica: uma introdução ao software R” há maiores explicações sobre o software e sobre sua utilização.

3. População e Amostra

Alguns conceitos:

População: é o conjunto de todos os indivíduos que se deseja estudar.

Amostra: é o subconjunto que será estudado formado pela seleção de indivíduos de determinada população, ou seja, os indivíduos nos quais são medidas ou observadas as características de interesse.

Exercícios:

3.1) Para cada alternativa abaixo, escreva se no estudo foi utilizado uma amostra ou população.

a) Um estudante de graduação realiza um projeto de pesquisa sobre os hábitos de alimentação dos moradores de Belo Horizonte. Ele verificou que apenas 25% das 278 pessoas selecionadas aleatoriamente consomem vegetais regularmente.

É amostra, pois foram selecionadas apenas alguns moradores da cidade de Belo Horizonte.

b) Foi realizada uma análise microbiológica na Faculdade de Farmácia com 20 queijos minas frescal vendidos no Mercado Central de Belo Horizonte. O resultado foi que a maioria dos produtos continha níveis de *Listeria monocytogenes* superiores ao permitido pela ANVISA.

É amostra, porque foram utilizados apenas alguns dos queijos minas frescal que são comercializados no local.

c) Uma nutricionista, a fim de investigar as fraudes de azeites comercializados na sua cidade, coletou 10 desses produtos e percebeu que menos de 1% deles foram fraudados.

É amostra, porque foram utilizados apenas alguns dos azeites comercializados na cidade.

d) Um educador físico enviou um questionário para os 50 clientes de seu Crossfit para que respondessem à pergunta: “Você já se lesionou durante alguma prática física fora do Crossfit?”.

Se todos responderam é população, se só alguns é amostra.

e) Um pesquisador realizou um estudo sobre o crescimento dos pés de pequi no Cerrado mineiro. Ele analisou 1200 pequizeiros em várias regiões do estado. Foram medidas a largura e a altura do tronco da árvore. Concluiu-se que os pequizeiros em Minas Gerais são menores que a média dos pequizeiros do Cerrado brasileiro.

É amostra, porque foram avaliadas apenas uma pequena seleção aleatória de árvores de pequi.

4. Tipos de variáveis

Alguns conceitos:

Variável: característica de interesse que é medida em cada indivíduo da amostra ou da população. Como o nome diz, seus valores variam de indivíduo para indivíduo.

Variáveis Quantitativas: são as características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que fazem sentido. Podem ser contínuas ou discretas.

- Variáveis contínuas: são características mensuráveis que assumem valores em uma escala contínua, em que quaisquer valores não-inteiros (com casas decimais) em um intervalo fazem sentido. Normalmente são medidas com o auxílio de um instrumento.
  Exemplos: peso (balança), altura (régua), índice pluviométrico (pluviômetro).

- Variáveis discretas: são características mensuráveis que podem assumir apenas um número contável de valores, nos quais só fazem sentido valores inteiros. Geralmente, são o resultado de contagens.
  Exemplos: número de filhos, número de telefones em uma loja, número de pessoas em uma casa.

Variáveis Qualitativas (ou categóricas): são as características que não possuem valores quantitativos. São definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais.

- Variáveis nominais: não existe ordenação entre as categorias.
  Exemplos: sexo, cor dos olhos, fumante/não fumante, doente/sadio.

- Variáveis ordinais: existe uma ordenação entre as categorias.
  Exemplos: escolaridade (1°, 2°, 3° graus), estágio da doença (inicial, intermediário, terminal).

  Observação: Algumas vezes, códigos numéricos podem ser atribuídos às possíveis respostas de uma variável qualitativa. Por exemplo, com a gravidade de uma doença, que ao invés de utilizarmos leve, moderado e grave, poderíamos utilizar uma escala de 1 a 3. Isso se torna necessário pois simplifica o uso de uma variável qualitativa como fator nos softwares estatísticos.

Exercícios:

4.1) Classifique as alternativas abaixo como sendo verdadeiras ou falsas. Para as falsas, justifique sua escolha e corrija a alternativa de modo que ela fique correta.

a) O tipo sanguíneo (A, B, AB e O) é uma variável qualitativa ordinal.

Falso: é uma variável qualitativa nominal, porque expressa valores não numéricos sem uma ordem hierárquica.

O tipo sanguíneo (A, B, AB e O) é uma variável qualitativa nominal.

b) O número de aparelhos eletrônicos por indivíduo de uma cidade é uma variável quantitativa discreta.

Verdadeiro

c) O volume de água (em litros) que uma mulher adulta consome diariamente é uma variável quantitativa contínua.

Verdadeiro

d) O grau de proficiência em inglês (A1, A2, B1, B2, C1) é uma variável qualitativa nominal.

Falso: é uma variável qualitativa ordinal porque seus valores são numéricos e apresentam uma hierarquia entre si: o nível C1 > B2 > B1 > A2 > A1

O grau de proficiência em inglês (A1, A2, B1, B2 e C1) é uma variável qualitativa ordinal.

e) O número de frutos colhidos em uma fazenda produtora de lichia é uma variável quantitativa discreta.

Verdadeiro

f) O peso da refeição em um restaurante (em g) é uma variável quantitativa discreta.

Falso: é uma variável quantitativa contínua, porque os valores não inteiros fazem sentido e são mensurados por uma balança.

O peso da refeição em um restaurante (em g) é uma variável quantitativa contínua.

g) O pH de um tomate, medido pela escala de Kasvi, é uma variável quantitativa discreta.

Verdadeiro

h) Frequência de idas à academia (diária, semanal ou quinzenal) é uma variável quantitativa ordinal.

Verdadeiro

i) A circunferência da cintura de mulheres (em cm) é uma variável quantitativa discreta.

Falso: é uma variável quantitativa contínua, porque os valores não inteiros fazem sentido e são mensurados por uma fita métrica.

A circunferência da cintura de mulheres (em cm) é uma variável quantitativa contínua.

j) A circunferência da cintura de mulheres classificada como baixa, alta e muito alta é uma variável quantitativa contínua.

Falso: É uma variável qualitativa ordinal, porque não apresenta valores numéricos e tem uma ordem hierárquica dada por: muito alta > alta > baixa.

A circunferência da cintura de mulheres classificada como baixa, alta e muito alta é uma variável qualitativa ordinal.

5. Distribuição de Frequências e Tabelas de Contingência

Frequência absoluta: é o número de vezes em que determinada observação apareceu na amostra. Uma tabela de frequências contém as informações a respeito das frequências dos diferentes níveis de uma variável.

Frequência relativa: é a divisão da frequência absoluta pelo número total de observações. Como são dados percentuais, permite uma melhor visualização e comparação entre diferentes grupos.

Frequência acumulada: é a soma dos valores anteriores até o valor atual, podendo se tratar da frequência absoluta ou da relativa. A última frequência acumulada será o total de observações ou então 100% das observações, quando se tratar de frequência relativa.

Tabela de contingência: são tabelas formuladas com o intuito de apresentar duas ou mais variáveis, podendo assim sintetizar melhor os dados, ou até mesmo analisar relação entre as variáveis.

Exercícios:

5.1) Foram coletados dados do ambulatório médico do Hospital das Clínicas referentes ao nível de dor (em uma escala de 1 a 5) de mulheres adultas com enxaqueca. Os dados estão expressos na tabela abaixo.

Tabela do nível de dor de enxaqueca de mulheres adultas	
Nível da dor	**Frequência Absoluta**
1	30
2	36
3	10
4	60
5	24
Total	160

a) Calcule a frequência relativa para o nível da dor encontrada neste grupo.

nível da dor 3: $f_3 = \frac{10}{160} \times 100\% = 6,25\%$

nível da dor 4: $f_4 = \frac{60}{160} \times 100\% = 37,50\%$

Para os outros níveis de dor, utilize a mesma lógica.

Tabela do nível de dor de enxaqueca de mulheres adultas		
Nível da dor	**Frequência Absoluta**	**Frequência Relativa**
1	30	18,75%
2	36	22,50%
3	10	6,25%
4	60	37,50%
5	24	15,00%
Total	160	100%

b) Interprete os dados.

Dentre as 160 mulheres do ambulatório do Hospital das Clínicas que apresentavam enxaqueca, a maior parte delas relatava nível de dor maior ou igual a 3. Demonstrando, assim, que as mulheres adultas que buscam atendimento ambulatorial nesse hospital normalmente apresentam níveis de dor mais altos. Além disso, pode-se observar que o nível de dor menos relatado foi o 3.

5.2) Em uma investigação dos fatores de risco para as doenças cardiovasculares, os níveis séricos de cotinina (produto metabólico da nicotina) foram registrados para um grupo de fumantes e um grupo de não fumantes. As distribuições de frequências correspondentes são mostradas abaixo.

Tabela de níveis séricos de cotinina de fumantes e não fumante

Nível de cotinina (mg/ml)	Fumantes	Não fumantes
0-13	78	3300
14-49	133	72
50-99	142	23
100-149	206	15
150-199	197	7
200-249	220	8
250-259	151	9
300-399	412	11
Total	1539	3445

a) É correto comparar as distribuições dos níveis de cotinina para fumantes e não fumantes, com base nas frequências absolutas em cada intervalo? Por que?

Não é correto comparar a distribuição de frequência absoluta dos níveis de cotinina para fumantes e não fumantes, porque o número de indivíduos por grupo (fumantes e não fumantes) não é o mesmo. Sendo assim, o adequado seria comparar conforme a frequência relativa de cada um.

b) Caso sua resposta para o item “a” seja negativa, construa uma nova tabela, em que as distribuições dos níveis de cotinina para fumantes e não fumantes possam ser comparadas.

Tabela ajustada de níveis séricos de cotinina de fumantes e não fumante

Nível de cotinina (mg/ml)	*Fumantes*	*Freq. relativa (Fumantes)*	*Não fumantes*	*Freq. relativa (Não Fumantes)*
0-13	*78*	*5,06%*	*3300*	*95,79%*

14-49	133	8,64%	72	2,09%
50-99	142	9,22%	23	0,67%
100-149	206	13,38%	15	0,43%
150-199	197	12,80%	7	0,20%
200-249	220	14,29%	8	0,23%
250-259	151	9,81%	9	0,26%
300-399	412	26,77%	11	0,31%
Total	1539	100%	3445	100%

c) Qual é a classe modal do nível de cotinina para não fumantes? O que isso significa?

A classe modal do nível de cotinina para não fumantes é 0-13. Isso significa que cerca de 96% dos não fumantes apresentam o menor nível de cotinina possível (0-13 mg/ml).

d) Analise as tabelas e suas respostas aos itens b e c. O que você pode dizer sobre a distribuição dos níveis de cotinina registrados para cada grupo?

No grupo de fumantes, o nível de cotinina sanguíneo varia mais, ou seja, a distribuição é menos homogênea. Já para fumantes, a distribuição está mais concentrada em baixos níveis sanguíneos de cotinina. Dessa forma, há uma relação entre maior nível de cotinina sanguínea e o hábito de fumar; o que já é esperado, visto que a cotinina é um produto metabólico da nicotina (substância presente no cigarro).

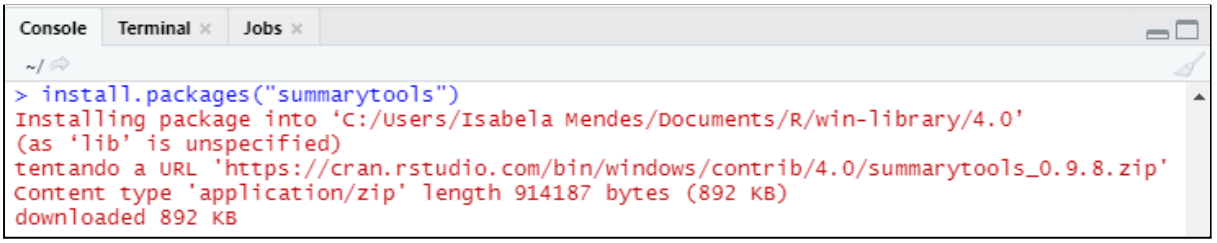
e) Para todos os indivíduos nesse estudo, o status do fumo é auto-registrado. Você acha que algum dos indivíduos pode estar mal classificado? Por que?

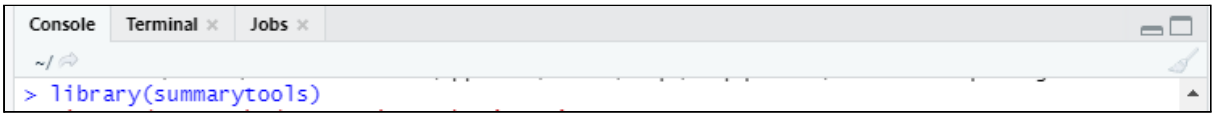
Alguns indivíduos podem estar mal classificados porque embora afirmem que não fumam, o nível de cotinina no sangue é significativamente alto. Como a cotinina é produto da nicotina, e a maior fonte de nicotina é o cigarro, essa concentração sanguínea alta de cotinina possivelmente se relaciona ao hábito de fumar não relatado pelo participante. Mas vale dizer também, que esses indivíduos que afirmam não ser fumantes e apresentam elevada cotinina sérica podem ser fumantes passivos, ou seja, os indivíduos realmente não fumam mas estão constantemente em contato com a fumaça do cigarro, que é rica em nicotina.

Exercícios utilizando o R:

Com intuito de apresentar tabelas de frequência mais organizadas no software R, optou-se pela utilização de um pacote adicional, chamado “summarytools”. Para instalá-lo, basta utilizar o comando “install.packages("summarytools")” e em seguida carregar este pacote com o comando “library(summarytools)”:

Instalação do pacote SummaryTools:



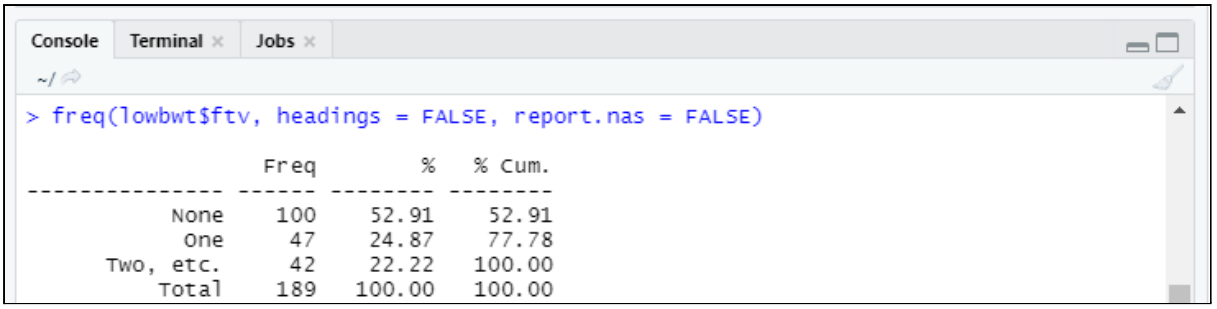


Para os exercícios que apresentarem entre parênteses um banco de dados e pacote, a resolução é realizada computacionalmente no software R utilizando este banco.

5.3) **(Banco de dados lowbwt - Pacote Aplore 3)** Se quisermos conhecer a amostra da pesquisa e relacionar o número de consultas com o peso de nascimento dos bebês, podemos utilizar tabelas de frequência. Sendo assim, crie uma tabela de frequência contendo frequência absoluta e relativa da variável de números de consultas médicas no primeiro semestre (ftv). Em seguida, faça uma tabela de contingência e observe se existe alguma relação com o baixo peso no nascimento dos filhos (low).

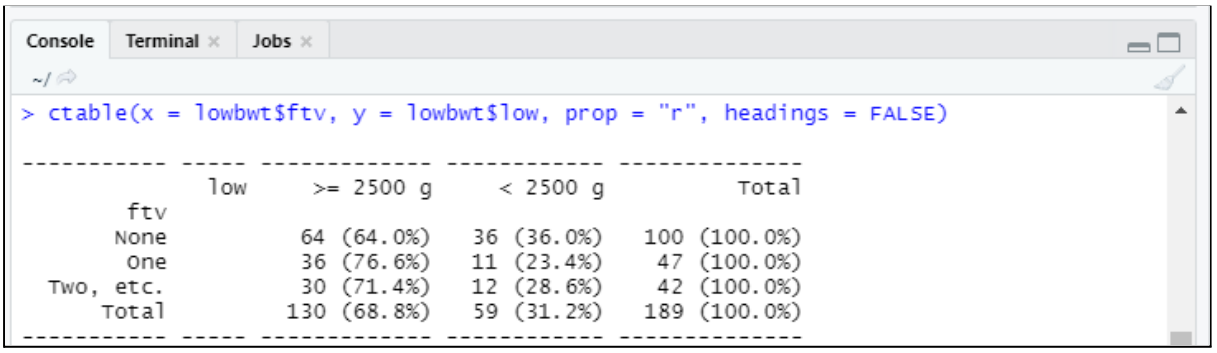
Para resolver essa questão no R vamos utilizar a seguinte função:

"freq" será a função que designa uma tabela de frequência; o símbolo "$" serve para selecionar qual variável do banco utilizar; para deixar a tabela com aspecto mais "limpo" usamos o argumento "headings = FALSE", que serve para retirar uma legenda que gera automaticamente, mas que não é necessária; já o comando "report.nas = FALSE" é utilizado para retirar duas colunas adicionais da tabela, as quais possuem a função de recalcular os valores excluindo participantes que não apresentaram o dado da variável em questão, como não possui dados faltando nessa variável, teríamos colunas com valores repetidos.



Interpretação: Com a tabela, podemos observar, por exemplo, que mais de 50% das mães não realizaram nenhuma consulta no primeiro trimestre de gestação, quase 80% realizaram nenhuma ou apenas 1 consulta no primeiro trimestre e que o número de mães que realizaram 1, 2 ou mais consultas é bem parecido.

Para a tabela de contingência, usamos a função "ctable" do pacote summarytools; em seguida, para indicar qual variável vai estar no eixo horizontal, utilize "x = lowbwt$ftv" e para o eixo vertical "y = lowbwt$low" ; o argumento " prop = "r" " vai indicar que o percentual entre parênteses é referente a categoria de sua linha horizontal. (Obs: para ter um percentual referente à categoria de sua coluna utiliza "t" ao invés de "r").



Interpretação: Com a tabela de contingência entre número de consultas no primeiro trimestre e peso do nascimento, podemos dizer que, nessa amostra, o percentual dos pesos de nascimento maiores ou iguais a 2500g é maior nos filhos das mães que fizeram pelo menos 1 consulta no primeiro trimestre, e que a proporção de bebês com mais do que 2500g ao nascer não foi maior no grupo de mães que fez 2 ou mais consultas .

5.4) **(Banco de dados lowbwt - Pacote Aplore 3)** Um estudo bem desenhado deve fazer o possível para evitar vieses em seus resultados. Por exemplo, uma pesquisa que avalia recuperação de cirurgias ortopédicas comparando dois grupos deve ter grupos com idades semelhantes, pois indivíduos mais novos provavelmente vão ter uma melhor recuperação, o que afetaria os resultados. Com isso, se torna importante ter uma amostra homogênea. Para saber se o banco de dados "lowbwt" possui uma amostra homogênea da variável peso da mãe, crie uma tabela de frequências absoluta, relativa e acumulada para a variável contínua peso da mãe no último período menstrual (lwt).

Para construir uma tabela de frequências para uma variável contínua, inicialmente temos que criar os intervalos para esta variável e, no software R, fazemos isso utilizando o comando abaixo:

Através do "categorizada" nomeamos a nova variável criada; o argumento "cut" será necessário para categorizar uma variável; em seguido "breaks" indicará quais são os limites dos intervalos; para dar nomes aos intervalos usamos o argumento "labels"; e por fim "c" combina valores em um vetor (conjunto de valores numéricos sob um único nome).

```
Console  Terminal  Jobs
~/
> categorizada<-cut(80:250, breaks = c(79,123,166,209,250), labels = c("80-123","124-1
66","167-209","210-250"))
> categorizada
  [1] 80-123  80-123  80-123  80-123  80-123  80-123  80-123  80-123  80-123
 [10] 80-123  80-123  80-123  80-123  80-123  80-123  80-123  80-123  80-123
 [19] 80-123  80-123  80-123  80-123  80-123  80-123  80-123  80-123  80-123
 [28] 80-123  80-123  80-123  80-123  80-123  80-123  80-123  80-123  80-123
 [37] 80-123  80-123  80-123  80-123  80-123  80-123  80-123  80-123  124-166
 [46] 124-166 124-166 124-166 124-166 124-166 124-166 124-166 124-166 124-166
 [55] 124-166 124-166 124-166 124-166 124-166 124-166 124-166 124-166 124-166
 [64] 124-166 124-166 124-166 124-166 124-166 124-166 124-166 124-166 124-166
 [73] 124-166 124-166 124-166 124-166 124-166 124-166 124-166 124-166 124-166
 [82] 124-166 124-166 124-166 124-166 124-166 124-166 167-209 167-209 167-209
 [91] 167-209 167-209 167-209 167-209 167-209 167-209 167-209 167-209 167-209
[100] 167-209 167-209 167-209 167-209 167-209 167-209 167-209 167-209 167-209
[109] 167-209 167-209 167-209 167-209 167-209 167-209 167-209 167-209 167-209
[118] 167-209 167-209 167-209 167-209 167-209 167-209 167-209 167-209 167-209
[127] 167-209 167-209 167-209 167-209 210-250 210-250 210-250 210-250 210-250
[136] 210-250 210-250 210-250 210-250 210-250 210-250 210-250 210-250 210-250
[145] 210-250 210-250 210-250 210-250 210-250 210-250 210-250 210-250 210-250
[154] 210-250 210-250 210-250 210-250 210-250 210-250 210-250 210-250 210-250
[163] 210-250 210-250 210-250 210-250 210-250 210-250 210-250 210-250 210-250
Levels: 80-123 124-166 167-209 210-250
```

Em seguida, podemos criar a tabela de frequência, utilizando a função "freq", também do pacote "summarytools":

```
Console  Terminal  Jobs
~/
> freq(categorizada, headings = FALSE, report.nas = FALSE)

                Freq        %   % Cum.
-------------- ------ -------- --------
       80-123     44    25.73    25.73
      124-166     43    25.15    50.88
      167-209     43    25.15    76.02
      210-250     41    23.98   100.00
        Total    171   100.00   100.00
```

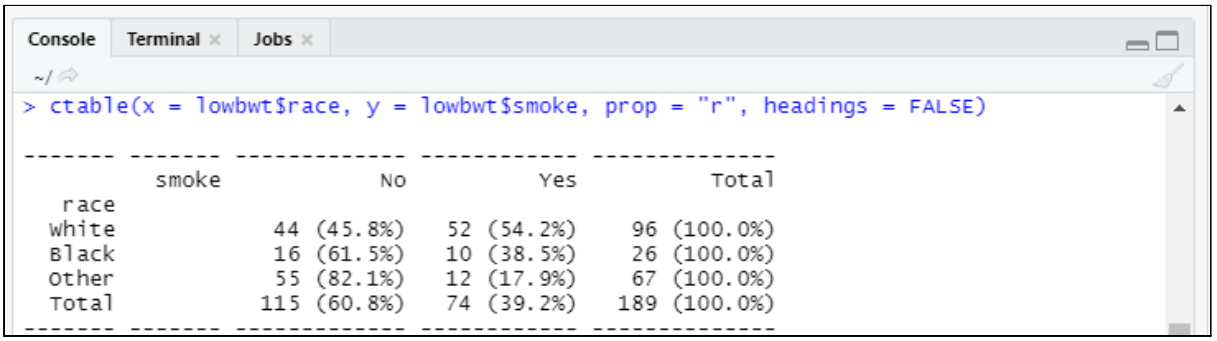
Interpretação: observa-se que, separada por intervalos aproximadamente iguais, a distribuição dos valores do peso na mãe no último período menstrual nessa amostra é bastante homogênea, o que contribui para conclusões menos enviesadas.

5.5) **(Utilizar banco lowbwt- Pacote Aplore 3)** Supondo que os pesquisadores querem realizar um estudo sobre gestantes, avaliando exposição a riscos e consequências,

descubra com base nessa amostra alguma característica que poderia ser de interesse para a pesquisa.

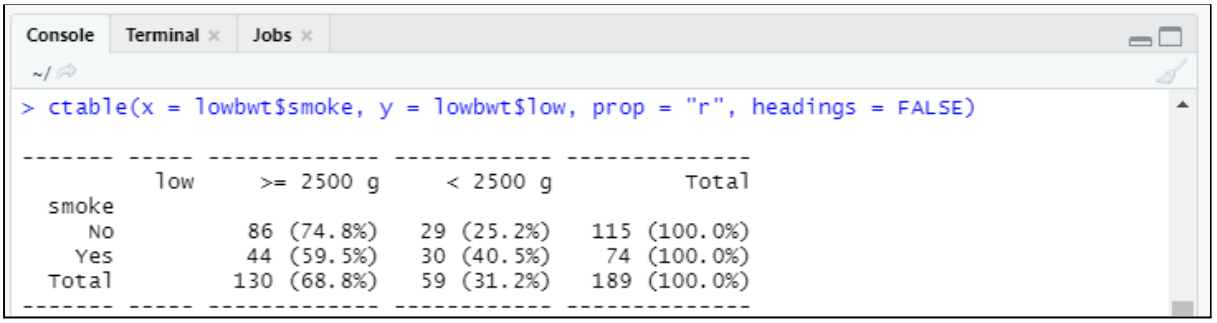
Para isso, poderia ser feito uma tabela de contingência para as variáveis etnia (race), e hábito de fumar (smoke), o que daria algumas informações a respeito das características e exposição a risco. E, para consequências, poderíamos criar uma tabela de contingência para as variáveis smoke e low, descobrindo se essa exposição afeta o peso do bebê.

Veja abaixo a resolução no software R, para a tabela de contingência com as variáveis race e smoke:



Interpretação: Observa-se que o hábito de fumar durante a gravidez foi maior em mulheres brancas (mais de 50%), em seguida em mulheres negras e em menor percentual nas demais etnias.

Agora veja a resolução no R para as variáveis smoke e low:



Interpretação: Já com relação ao hábito de fumar durante a gravidez e ao peso do nascimento, observa-se que houve um maior percentual de nascimento com peso ≥ 2500g nas mães que não fumaram, consequentemente, um maior percentual de nascimento <2500g em mães que fumaram durante a gravidez.

6. Medidas de tendência central

Média: é o valor que demonstra a concentração dos dados de uma distribuição, ou seja, o ponto de equilíbrio das frequências em um histograma.

Moda: é o valor mais frequente da variável, o que mais se repete.

Mediana: é o valor localizado na região central de um conjunto de dados ordenado, não necessariamente sendo um valor pertencente ao banco de dados.

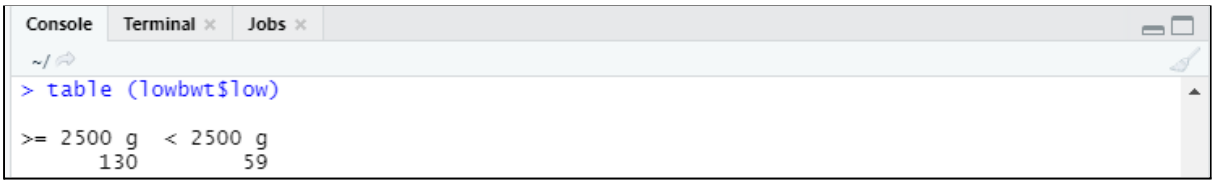
Quartil: O primeiro, segundo e terceiro quartis são medidas que dividem a amostra em quatro partes, contendo pelo menos 25% dos dados em cada uma destas partes. Por exemplo, a primeira parte é o subconjunto de dados com valores contidos entre o mínimo da amostra e o primeiro quartil. Os quartis são, respectivamente, os percentis de ordem 25%, 50% e 75%.

Exercícios:

6.1) **(Banco de dados lowbwt - Pacote Aplore3)** Foram coletados dados de 189 crianças de uma maternidade.

a) Calcule a frequência absoluta de crianças com baixo peso ao nascer e crianças com peso normal.

Utilizando o comando "table" obtemos a frequência absoluta de uma variável. Como a nossa variável de interesse é o peso dessas crianças, utilizamos o seguinte argumento: "lowbwt$low". Veja abaixo os comandos:



Interpretação: Dentre as crianças estudadas, 130 nasceram com peso maior ou igual a 2500g, enquanto 59 nasceram com peso inferior a 2500g. Sendo assim, a maioria das crianças nessa maternidade nasceram com peso considerado normal.

b) Qual a porcentagem de crianças com baixo peso ao nascer?

Para construir uma tabela com a frequência relativa do peso dessas crianças, inicialmente criamos um vetor a partir do comando "Tab=table(lowbwt$low)" e assim, salvamos a tabela em Tab. Em seguida utilizamos o comando "prop.table" para obtermos a frequência relativa do peso dessas crianças. Veja abaixo a resolução no software R:

```
Console  Terminal  Jobs
~/
> Tab=table(lowbwt$low)
> prop.table(Tab)

>= 2500 g  < 2500 g
0.6878307 0.3121693
```

A porcentagem de crianças com baixo peso ao nascer é 31,21%.
Interpretação: Como já mencionado anteriormente, menos da metade das crianças (31,21%) dessa maternidade nasceram com baixo peso.

c) Encontre a distribuição de frequências das idades das mães das crianças estudadas. Para isso, construa uma tabela utilizando intervalos com tamanho de 10 anos.

Para resolver essa questão pelo software R, definimos um vetor nomeado como "categorizada" através da função "cut", que define a qual intervalo cada valor da variável "lowbwt$age" pertence. O argumento "breaks" indica quais são os limites dos intervalos construídos. Por fim, a função "table" cria uma tabela de frequências por intervalos. Veja a resolução abaixo:

```
Console  Terminal  Jobs
~/
> categorizada<-cut(lowbwt$age, breaks = c(10,20,30,40,50))
> table(categorizada)
categorizada
(10,20] (20,30] (30,40] (40,50]
     69     100      19       1
```

Interpretação: Dentre as mães das crianças contempladas por esse estudo, a maioria possui idade entre 20 e 30 anos, sendo que apenas 1 tem mais de 40 anos. Podemos notar também que, nesse estudo, há uma concentração maior de mães com até 30 anos de idade.

d) Encontre a distribuição de frequência absoluta de peso dessas crianças. Para isso, construa uma tabela com 5 intervalos (ex: 500 a 1000g, 1000 a 2000g, …, 4000 a 5000g).

Para a resolução desse exercício utilizando o software R, criamos um vetor de nome "categorizada" através da função "cut" . Em seguida, definimos os valores mínimos e máximo encontrados pelo estudo, que são "709:4990", por fim, pelos argumentos "breaks" criamos os limites dos 5 intervalos sugeridos pelo enunciado. Feito isso, pela função "table" obtemos a tabela intervalada de frequência absoluta de peso das crianças deste estudo.

```
> categorizada<-cut(lowbwt$bwt,709:4990,breaks = c(500,1000,2000,3000,4000,5000), labels=
c("500-1000","1000-2000","2000-3000","3000-4000","4000-5000"))
> table (categorizada)
categorizada
 500-1000 1000-2000 2000-3000 3000-4000 4000-5000
        1        18        78        83         9
```

Interpretação: Dentre as crianças estudadas, a maioria nasceu com peso entre 2000g e 4000g.

e) Qual o peso médio ao nascer dessas crianças?

Resolvemos essa questão utilizando a função "mean" para obter a média, cujo argumento é "lowbwt$bwt", ou seja, a variável de interesse.

```
> mean(lowbwt$bwt)
[1] 2944.656
```

Interpretação: A média de peso dessas crianças era de 2944.656g, sendo, considerada, portanto, um peso médio adequado (> 2500g).

f) Qual a moda do peso dessas crianças?

Para isso, basta olhar nos dados qual o peso que mais se repete. Nesse caso, 4 pesos têm a mesma repetição, com 4 crianças cada. São eles: 2495g; 2920g; 2977g e 3651g. Sendo assim, a moda é um conjunto com valores 2495g; 2920g; 2977g e 3651g.

g) Calcule a mediana do peso dessas crianças.

Pela função "mean" obtemos a mediana cujo argumento "Lowbwt$bwt" é a variável de interesse. Observe abaixo:

```
> median(lowbwt$bwt)
[1] 2977
```

Interpretação: A mediana, ou seja, o número do meio dessa sequência, é 2977g.

6.2) (Fuvest 93 - Adaptada) A distribuição dos salários dos funcionários de um hospital é dada na tabela a seguir:

Tabela de distribuição de salários dos funcionários do hospital

Salário (em Cr$)	Nº de funcionários
500.000,00	10
1.000.000,00	5
1.500.000,00	1
2.000.000,00	10
5.000.000,00	4
10.500.000,00	1
Total	31

a) Qual é a média e qual é a mediana dos salários desse hospital?

$$\overline{x} = \frac{(500\times10)+(1.000.000\times5)+(1.500.000\times1)+(2.000.000\times10)+(5.000.000\times4)+(10.500.000\times1)}{31}$$

=2.000.000

Md= 1.500.000

Para encontrar a mediana é só imaginar os valores ordenados de forma crescente, e pegar o valor do meio, como temos um total de 31 funcionários o salário mediano será o do respectivo funcionário que representará o valor 16 na ordem crescente, como a tabela já está em ordem crescente facilita a observação.

b) Suponha que sejam contratados dois novos funcionários com salários de Cr$2.000.000,00 cada. A variância da nova distribuição de salários ficará menor, igual ou maior que a anterior?

Vai diminuir pois 2.000.000 é o mesmo valor da média, pensando no cálculo da variância esse valor não aumentará a soma dos quadrados, porém irá aumentar duas unidades no denominador, diminuindo o resultado da divisão.

6.3) Os números mensais de internados em um hospital nos últimos 6 meses foram respectivamente de 84; 91; 72; 68; 87 e 72. O número médio, mediano e modal de internados nesse hospital é de respectivamente:

a) 79; 78; 72

b) 72; 78; 79

c) 78; 78; 79

d) 72; 78; 79

e) 78; 79; 72

$\overline{x} = \frac{84+91+72+68+87+72}{6} = 79$

Ordenando os valores em forma crescente temos: 68; 72; 72; 84; 87; 91
Temos nesse caso dois valores centrais 72 e 84 , nesse caso a mediana será a média entre esses números:

$\frac{72+84}{2} = 78$

Já a modal é o valor que mais se repete, nesse caso é o valor 72. Portanto temos média 79; mediana 78; e modal 72, Alternativa A

6.4) (G1 v4.11) (FUVEST/G.V.92) Num determinado país a população feminina representa 51% da população total. Sabendo-se que a idade média (média aritmética das idades) da população feminina é de 38 anos e a da masculina é de 36 anos. Qual a idade média da população?

a) 37,02 anos

b) 37,00 anos

c) 37,20 anos

d) 36,60 anos

e) 37,05 anos

Atenção nesse caso pode parecer que precisamos apenas fazer a média entre 36 e 38, no entanto se trata de uma média ponderada, devemos considerar a proporção de mulheres e homens, para facilitar podemos utilizar a frequência relativa, portando 0,51 para mulheres e 0,49 para os homens.

$\overline{x} = (0,51 \times 38) + (0,49 \times 36) = 32,02$

6.5) (Unb 99 - Adaptada) A tabela adiante apresenta o levantamento do número de faltas a cada 100 exames agendados em uma determinada clínica, durante um período de 30 dias úteis.

Tabela de faltas aos exames a cada 100 agendados durante 30 dias

Dia	Nº de faltas	Dia	Nº de faltas	Dia	Nº de faltas
1	6	11	1	21	2
2	4	12	5	22	6
3	3	13	4	23	3
4	4	14	1	24	5
5	2	15	3	25	2
6	4	16	7	26	1
7	3	17	5	27	3
8	5	18	6	28	2
9	1	19	4	29	5
10	2	20	3	30	7

Considerando que S é a série numérica de distribuição de frequências do número de faltas a cada 100 agendamentos, julgue os itens abaixo.

(a) A moda da série S é 5.

(b) Durante o período de levantamento desses dados, o percentual de faltas ficou, em média, abaixo de 3,7%.

(c) Os dados obtidos nos 10 primeiros dias do levantamento geram uma série numérica de distribuição de frequências com a mesma mediana da série S.

Para essa questão o ideal é que seja construído uma tabela de frequências, apesar de ser possível solucionar apenas contando os valores na tabela já presente, a organização dos dados é uma parte fundamental na estatística, e quando se trabalha com um maior número de dados essa organização se torna cada vez mais necessária.

Tabela de frequência para o número de faltas a cada 100 agendamentos

Número de faltas em 30 dias	Frequência absoluta (ni)	Frequência acumulada absoluta (Faci)
1	4	4
2	5	14
3	**6**	32
4	**5**	52
5	5	77
6	3	95
7	2	109
		Total=109

Afirmação (a) FALSA: a moda é o número que mais se repete entre todos os valores, nesse caso a moda é o valor 3

Afirmação (b) VERDADEIRA: Basta somar o número de faltas de todo o mês e dividir por 30, ou seja, 109/30=3,63 (média amostral)

Afirmação (c) VERDADEIRA: Como o conjunto de dados tem uma quantidade par de observações, bastas encontrar as observações que ocupam a 15° e 16° posições no banco de dados ordenado e calcular a média entre elas:

$\frac{3+4}{2}$ = 3,5

Para achar a mediana dos 10 primeiros dias podemos colocá-los em sequência, ou construir outra tabela caso se tratasse de uma maior quantidade de dados, como não é o caso...

1,2,2,3,**3**,**4**,4,4,5,6: portanto a mediana será:

$\frac{3+4}{2}$ = 3,5

Observação: Os valores grifados são referentes aos valores centrais utilizados para cálculo da mediana.

6.6) (Ufu 99 - Adaptada) Um pequeno hospital possui 30 funcionários com a seguinte distribuição salarial em reais.

Tabela de distribuição de salário entre os funcionários de um hospital

Nº de funcionários	Salários em R$
10	2.000,00
12	3.600,00
5	4.000,00
3	6.000,00

Quantos funcionários que recebem R$3.600,00 devem ser demitidos para que a mediana desta distribuição de salários seja de R$2.800,00?

a) 8

b) 11

c) 9

d) 10

e) 7

Para se obter uma mediana diferente dos valores presentes, pode se deduzir que a mediana será a média entre dois valores. Para se obter o valor de 2800 reais, a única possibilidade é fazendo a média de 2000+3600, ou seja, os valores centrais precisam ser esses, sabendo-se disso, como existe apenas 10 funcionários recebendo salário de 2000 reais, o 10º funcionários (ordenados em ordem crescente de salário) precisa estar em um dos valores centrais, o 11º deverá receber salário de 3600 reais, portanto vamos demitir a quantidade necessária para se obter um total de 20 funcionários (o que faz a mediana ser a média entre o 10º e o 11º), e assim obter mediana de 2800, ou seja, teremos que demitir 10 funcionários dos 12 que recebem 3600 reais. Alternativa D.

6.7) **(Banco de dados ICU - Pacote Aplore3)** Sabe-se que a expectativa de vida da população estudada é de 76 anos. Levando isso em consideração, descubra se os pacientes que vieram a óbito atingiram a expectativa de vida da população, em seguida faça

uma comparação entre a média de idade entre os pacientes que que vieram a óbito e dos que sobreviveram. O que se pode supor desses resultados?

Observação: a variável morte pode ser "died" ao invés de "sta", dependendo da sua versão do pacote.

Para responder essa questão podemos calcular a média de idade dos pacientes que morreram e dos que sobreviveram, para isso utilize o seguinte comando:

A função "by" estabelece uma separação de grupos para o resultado que temos; o argumento "mean" indica que queremos a média; colocando na seguinte ordem mostrada abaixo, isso determinará que vamos ter a média da idade, separada pelo desfecho da variável "died" a qual tem como resposta "yes" ou "no".

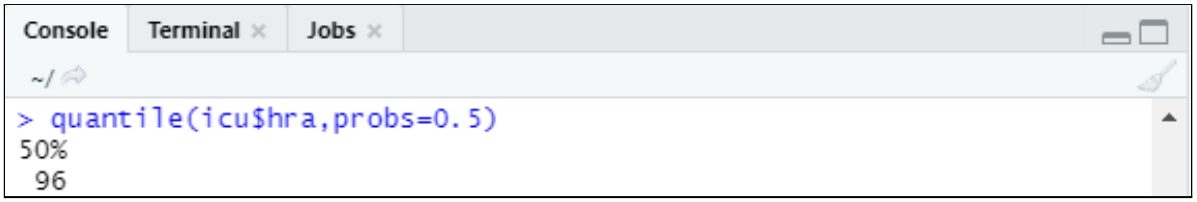
```
> by(ICU$age, ICU$died, mean)
ICU$died: No
[1] 55.65
------------------------------------
ICU$died: Yes
[1] 65.125
```

Interpretação: Além dos pacientes que vieram a óbito não terem alcançado a expectativa de vida de sua população, a idade se mostrou relevante para o desfecho morte, apresentando uma diferença de aproximadamente 10 anos na média de idade dos dois grupos.

6.8) **(Utilizar banco icu- Pacote Aplore3)** Considerando que um estudo anterior tenha aferido a frequência cardíaca de 101 indivíduos em leitos de internação, e que esses, quando ordenados de forma crescente, apresentaram o valor da 51º observação cardíaca igual a 80 bpm (note que este valor é igual a uma medida de tendência central), descubra utilizando a mesma medida de resumo, se a frequência cardíaca na admissão na UTI (variável hra), da amostra do banco ICU, é diferente da frequência de pacientes em leitos de internação do estudo anterior.

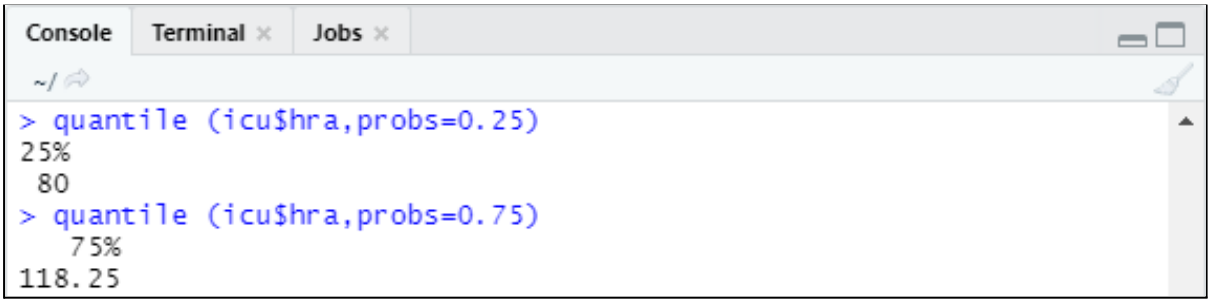
Vamos resolver encontrando o segundo quartil, ou percentil 50, que é o mesmo que a mediana. Por meio do R obtemos esses valores da seguinte forma:

A função "quantile" é usada para obter o valor de determinado quartil, no caso o segundo, visto que "probs" = 50%.



Interpretação: a mediana da frequência cardíaca dos pacientes em UTI nesta amostra foi de 96 bpm, o mesmo que o dizer que o segundo quartil é 96 bpm, ou seja, a frequência cardíaca desses pacientes se mostrou elevada em comparação a frequência cardíaca dos pacientes em leitos de internação do estudo anterior.

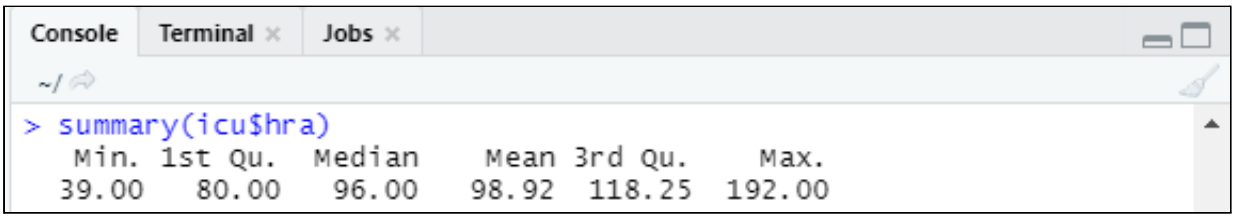
Observação: Perceba que o comando "probs = X" pode ser usado para encontrar os demais quartis. Como mostrado abaixo:



6.9) **(Banco de dados ICU - Pacote Aplore3)** Suponha que após a coleta de dados, o enfermeiro responsável pela ala de UTI te peça uma tabela com os dados resumidos sobre a frequência cardíaca desses pacientes, exigindo valor mínimo, máximo, 1º, 2º e 3º quartil e média. Qual seria uma forma rápida de apresentar todas essas medidas de resumo?

Para isso podemos pedir uma tabela resumida para a variável frequência cardíaca na admissão na UTI (hra).

Faça isso utilizando "summary", comando para obter essas principais medidas de resumo de uma dada variável.



Interpretação: Dentre os pacientes os valores mínimo e máximo foram respectivamente de 39 e 192 bpm; 50% dos pacientes apresentaram frequência cardíaca entre 80 a 118,25 bpm, com valor mediano de 96 bpm; a média geral da frequência cardíaca foi de 98,92 bpm.

7. Tipos de gráficos

Gráfico de barras: é aquele em que o tamanho de cada barra corresponde à frequência de uma das categorias.

Gráfico de setores: é aquele construído por um círculo, em que cada setor corresponde à frequência de determinada categoria.

Histograma: é um gráfico de barras construído com as barras unidas, devido ao caráter contínuo dos valores da variável

Boxplot (Diagrama de caixa): gráfico que utiliza uma caixa para representar os valores dos quartis, assim como linhas para expressar valores de limite máximo e mínimo, e pontos para representar valores discrepantes/atípicos. Além disso, ainda é possível comparar a simetria, tendência central e variabilidade entre diferentes grupos com base no formato obtido do Boxplot.

Diagrama/Gráfico de dispersão: indivíduos são representados com um ponto e marcados em um gráfico conforme seu valor da variável no eixo X e seu valor na variável do eixo Y, conforme o "desenho" formado pode-se estabelecer a existência ou não da relação entre essas variáveis, bem como a intensidade desta caso exista.

Gráfico de linha: Gráfico que apresenta uma variável de tempo no eixo X para acompanhar as mudanças ao longo de um determinado período sobre uma variável quantitativa, que será o eixo Y, formando ao final uma linha.

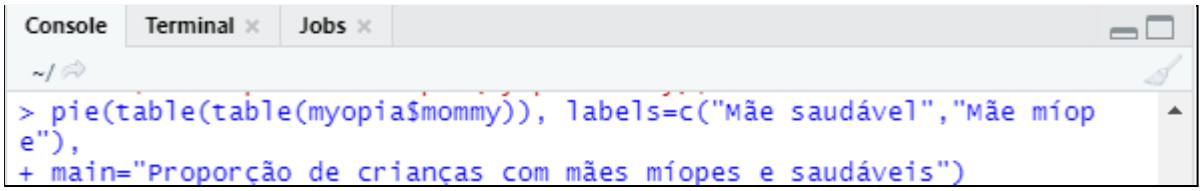
Exercícios:

7.1) **(Banco de dados myopia - Pacote Aplore 3)** Com os dados coletados de 618 crianças míopes, um pesquisador avaliou a relação entre a prevalência de miopia durante o período de 5 anos de acompanhamento.

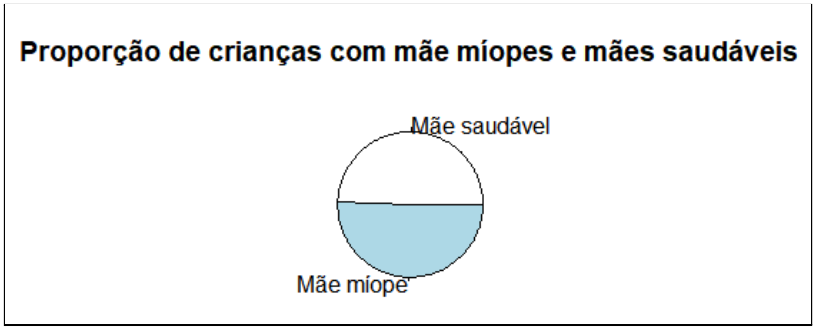
a) Apresente, em um gráfico de setores, a proporção de crianças do estudo que tinham mães míopes e não míopes. Interprete o resultado.

Aplicando no R, digitamos a função "pie" para criar um gráfico de setores, em seguida utilizamos "table" para obtermos a proporção de crianças com mães míopes e saudáveis. Para nomear cada setor, aplicamos o argumento "labels" e "main" para intitular o gráfico.

Observe a seguir os comandos no console do R:



Resultado:



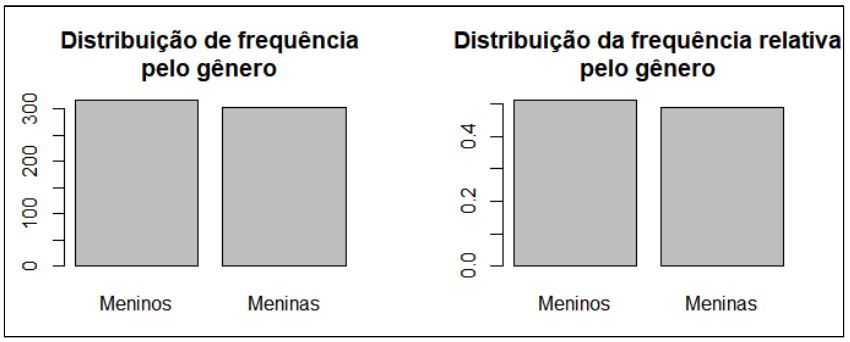
Interpretação: Embora a proporção de crianças com mães míopes seja maior que a proporção de crianças com mães não míopes, a diferença é bem pequena. Não podemos afirmar com precisão que não há relação entre a mãe ser míope e o filho ser míope porque não sabemos se as proporções de mães míopes e não míopes são aproximadamente iguais na população. Por exemplo, se a proporção de mães com miopia na população for de 20%, a afirmação de que a miopia do filho não está relacionada com a mãe só pode ser feita se for observada uma prevalência de 20% de miopia entre as mães desta amostra.

b) Faça dois gráficos de barras, com as distribuições absoluta e relativa, mostrando a distribuição de frequências do sexo biológico das crianças do estudo.

Ainda aplicando o R, a função "par(mfrow=c(1,2))" é utilizada para que se obtenha dois gráficos lado a lado. Em seguida, é usado "barplot" para criar um gráfico de barras; e os argumentos "names" nomeiam cada coluna, enquanto "main" intitula o gráfico da forma requerida.

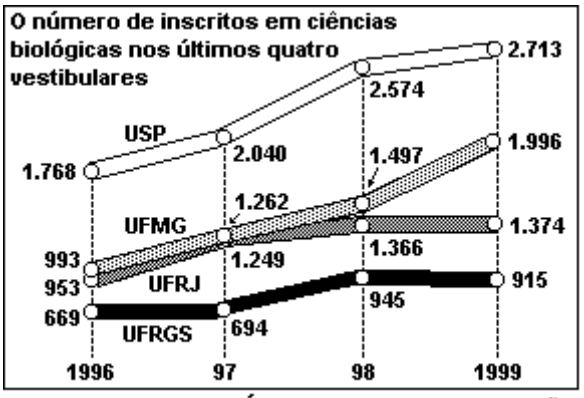
Observe o comando no R para frequência absoluta:

```
Console   Terminal   Jobs
~/
> par(mfrow=c(1,2))
> barplot(table(myopia$gender),names.arg=c("Meninos","Meninas"),
+ main= "Distribuição de frequência
+ pelo gênero")
> barplot(prop.table(table(myopia$gender)), names.arg= c("Menino
s","Meninas"),
+ main="Distribuição da frequência relativa
+ pelo gênero")
```



Interpretação: Ambos os gráficos mostram que o número de meninas é aproximadamente igual ao número de meninos e representam, respectivamente 51,13% e 48,86% das crianças do estudo.

7.2) (UnB 2000 - Adaptada) O gráfico a seguir ilustra o número de inscritos nos últimos quatro vestibulares que disputaram as vagas oferecidas pela Universidade de São Paulo (USP) e pelas universidades federais do Rio de Janeiro (UFRJ), de Minas Gerais (UFMG) e do Rio Grande do Sul (UFRGS).



Época, 26/4/99 (com adaptações).

Com base nessas informações, julgue os itens seguintes:

(a) De 1997 a 1998, o crescimento percentual do número de inscritos na USP foi maior que o da UFRGS.

(b) Todos os segmentos de reta apresentados no gráfico têm inclinação positiva.
(c) Durante todo período analisado, a UFMG foi a universidade que apresentou o maior crescimento percentual, mas não o maior crescimento absoluto.
(d) Os crescimentos percentuais anuais na UFRJ diminuíram a cada ano.
(e) Considerando, para cada universidade representada no gráfico, a série numérica formada pelos números de inscritos em ciências biológicas nos últimos quatro vestibulares, a série da USP é a que apresenta a maior mediana, tendo desvio-padrão maior que o da UFRJ.

(a) FALSA: Apesar de em número a USP de fato ter um aumento maior, a questão afirma sobre o percentual, o qual quando calculamos, o crescimento percentual da USP de 1997 a 1998 teve um aumento de 26,18%, e o crescimento percentual da UFRGS no mesmo período foi 36,17%.

(b) FALSA: Houve uma exceção, nos anos de 1998 à 1999 na UFRGS a reta apresentou segmento negativo, com queda no número de inscritos.

(c) FALSA: O percentual de aumento total nos 4 anos das universidades foi de 53,45% para a USP; de 101,01% para a UFMG; de 44,18% UFRJ; e de 36,77% para a UFRGS. Quanto ao crescimento absoluto, obteve-se os valores de 945 inscritos a mais na USP; 1003 na UFMG; 421 na UFRJ; e de 246 na UFRGS; ou seja, a UFMG teve o maior crescimento percentual e também absoluto na inscrição para prestar vestibular no curso de ciências biológicas.

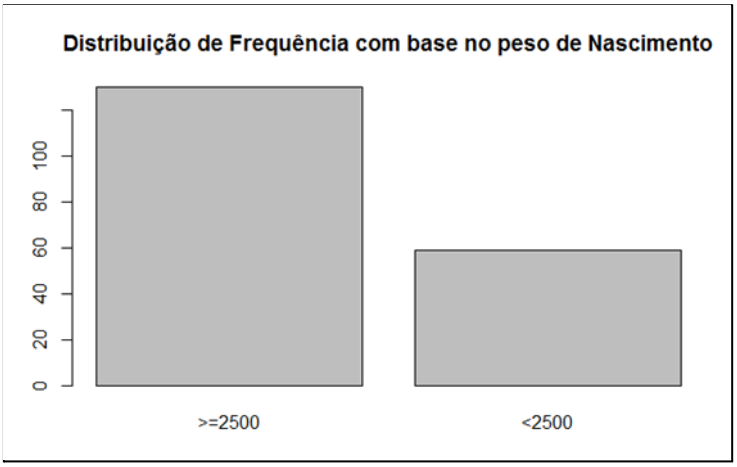
(d) VERDADEIRA: Nos anos de 1996 à 1997 o crescimento foi de 31,06%; já nos anos de 1997 à 1998 foi de 9,37%; e por fim nos anos de 1998 à 1999 foi de apenas 0,59%.

(e) VERDADEIRA: A mediana da USP é de 2307, já se sabe que será a maior mediana porque nenhuma outra universidade chegou a valores maiores que 2000 inscritos no vestibular. Também terá maior desvio-padrão pelo fato de que o desvio padrão é uma medida de dispersão, como o intervalo entre o número máximo e mínimo de inscritos na USP é maior, ou seja, mais disperso, consequentemente ele terá também um maior desvio padrão em comparação a UFRJ.

7.3) **(Utilizar banco lowbwt - Pacote Aplore 3)** O grupo que realizou a pesquisa da origem do banco de dados, deseja apresentar seus resultados em um congresso de ginecologia e obstetrícia. Para isso, querem montar um gráfico que apresente de forma visual a variável peso separada em dois grupos. Assim sendo, monte um gráfico de barras para a variável de baixo peso (low), com uma barra para peso maior ou igual a 2500g e outra para peso menor que 2500g.

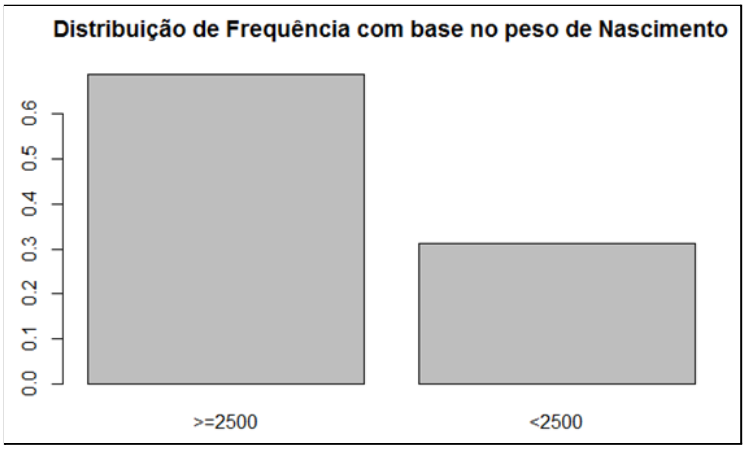
Para a resolução do exercício podemos utilizar os seguintes comandos para fazer um gráfico de barras considerando a frequência absoluta, veja os comandos a seguir:

```
> par(mfrow=c(1,2))
> barplot(table(lowbwt$low),names.arg = c(">=2500","2500"), main = "Distribuição
 de Frequência com base no Peso de Nascimento")
>
```



Note que o gráfico acima não permite uma comparação em termos percentuais. Para obtener um gráfico de barras com frequências relativas utilizamos os comandos abaixo:

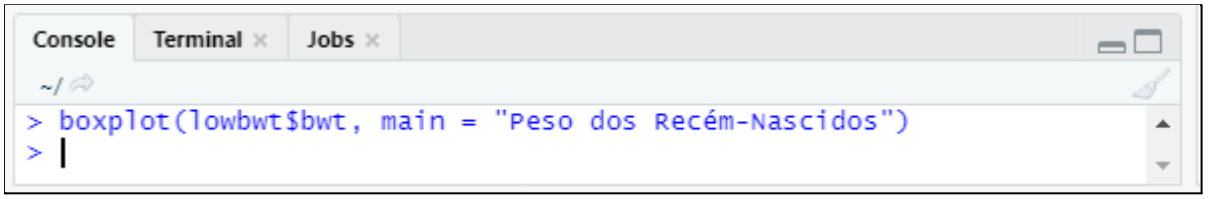
```
> par(mfrow=c(1,2))
> barplot(prop.table(table(lowbwt$low)),names.arg = c(">=2500","2500"), main =
 "Distribuição de Frequência com base no Peso de Nascimento")
>
```

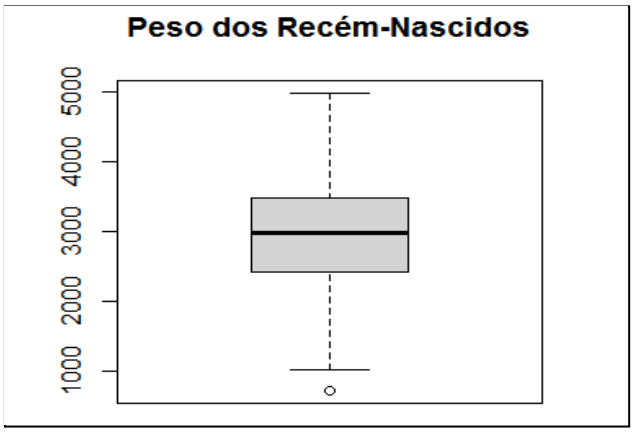


Interpretação: nesta amostra existem mais de 120 bebês com peso maior ou igual a 2500 gramas e em torno de 60 com peso menor que 2500 gramas. Em percentual, temos que 70% dos recém-nascidos têm peso maior que 2500 gramas e 30% menor que 2500 gramas. Portanto, a maior parte dos recém-nascidos nesse estudo apresentam peso adequado, se tomarmos como parâmetro adequado como maior ou igual a 2500g.

7.4) **(Banco de dados lowbwt - Pacote Aplore 3)** Para compreender melhor a distribuição do peso dos recém-nascidos, os pesquisadores querem descobrir se existem valores atípicos na amostra. Para isso construa um bloxpot para a variável peso observado no nascimento (bwt).

Para criação de um boxplot no R iremos utilizar a função "boxplot", em que informamos a variável utilizada e o título do gráfico.



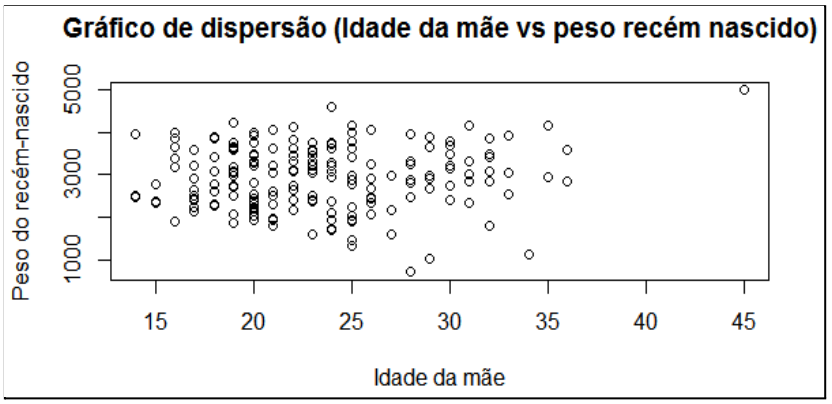


Interpretação: Observa-se a presença de um valor atípico abaixo de 1000g; o segundo quartil (Percentil 50) de peso em torno de 3000g, indica também o valor da mediana; o primeiro quartil (Percentil 25) em torno de 2500g e terceiro quartil (Percentil 75) em torno de 3500 g; assim sendo houve uma diferença aproximada de 500g entre o segundo quartil em relação ao primeiro e terceiro quartil, mostrando uma distribuição simétrica, e também uma baixa dispersão, quando observamos o tamanho da caixa em relação ao restante do gráfico.

7.5) **(Banco de dados lowbwt - Pacote Aplore3)** Construa um diagrama de dispersão para avaliar se existe alguma relação entre a idade das mães e os pesos dos recém-nascidos.

Por meio do R utilize o comando abaixo, no qual "plot" determinará que se trata de um gráfico de dispersão:

```
> library(aplore3)
> plot(lowbwt$age, lowbwt$bwt, main = "Gráfico de dispersão (Idade da mãe vs peso
 recém nascido)", ylab = "Peso do recém-nascido", xlab = "Idade da mãe")
>
```

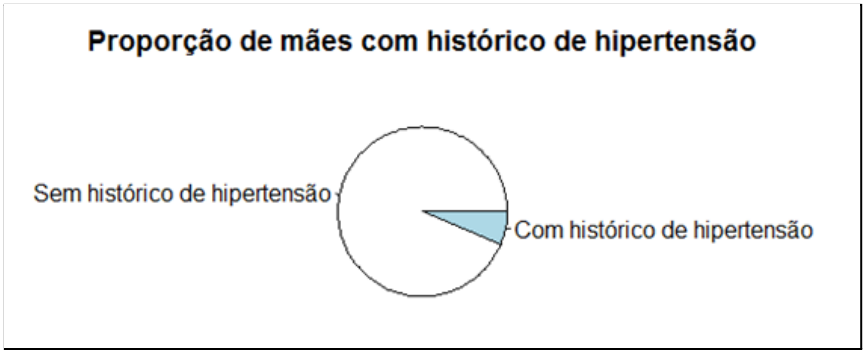


Interpretação: Pelo gráfico é possível ver que os pontos foram relativamente bem distribuídos, indicando não haver nenhuma relação entre essas duas variáveis.

7.6) **(Banco de dados lowbwt - Pacote Aplore3)** Sabendo-se que a hipertensão é um fator de risco para gestantes, conhecer o quanto das gestantes possuem esse risco gestacional é um ponto crucial durante a gestação. Sendo assim, construa um gráfico de pizza para a frequência de mães deste estudo que apresentam, ou não, histórico de hipertensão.

Operando no R novamente, utilize argumentos já descritos anteriormente:

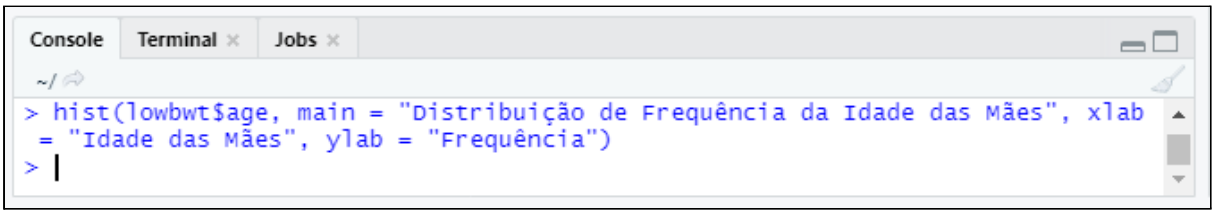
```
> pie(prop.table(table(lowbwt$ht)),labels = c("Sem histórico de hipertensão", "C
om histórico de hipertensão"), main = "Proporção de mães com histórico de hipert
ensão")
>
```

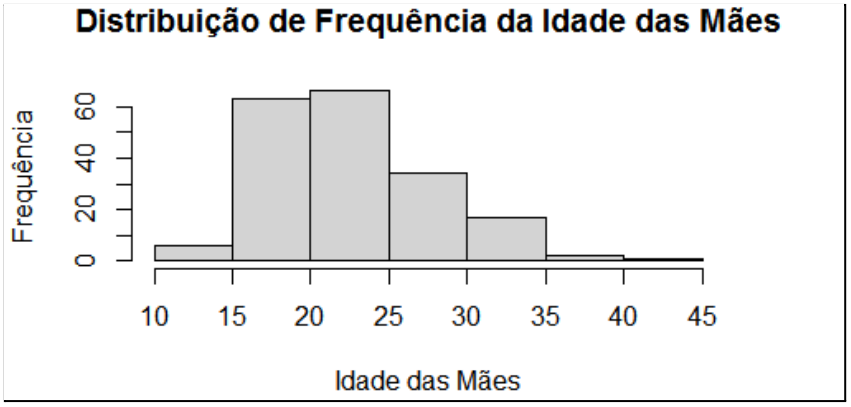


Interpretação: Com o gráfico de setores vemos que a maior parte das mães nessa amostra não possuía histórico de hipertensão. Sendo assim, uma minoria precisará de uma atenção maior no pré natal.

7.7) **(Bando de dados lowbwt - Pacote Aplore3)** Crie um gráfico para estudar a distribuição de frequências da idade das mães neste estudo.

Uma boa forma de ilustrar essas informações seria em um histograma, o qual pode ser feito pela função "hist", como segue abaixo:





Interpretação: Pelo histograma vemos que a maior frequência de mães está entre 20 a 25 anos, com mais de 60 mães; essa amostra apresentou alta frequência de mães muito jovens entre 15 a 20 anos; a frequência decai depois de 25 anos, e podemos concluir que essa amostra possui uma distribuição normal (curva em formato de sino).

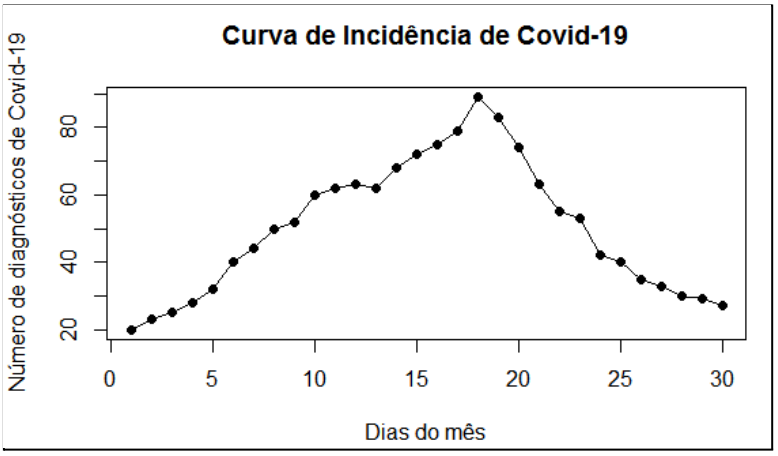
7.8) Durante a pandemia da COVID-19, em uma cidade percebeu-se que o número de casos estava aumentando. Assim, no dia 15 implementaram medidas para o seu controle. Para saber se as medidas foram eficazes no controle do pico de contágio, construa um gráfico de linha.

O gráfico de linha é uma excelente forma de avaliar variáveis temporais, para uma correta formulação deve-se colocar no eixo X a variável tempo e no eixo Y a variável com a frequência a ser analisada.

Para isso utilize o comando do gráfico de dispersão "plot", no entanto, apenas com essa função o gráfico não terá linhas conectando os pontos, portanto, para transformar em

um gráfico de linha, basta adicionar o argumento "type = "o"" que ligará os pontos entre si; além disso se preferir use "pch = 16" comando o qual preenche os pontos para um gráfico mais esteticamente agradável. Veja abaixo:

```
Console  Terminal  Jobs
~/
> plot(banco_gl$Data, banco_gl$Ndc19, main = "Curva de Incidência de Covid-19",
 xlab = "Dias do mês", ylab = "Número de diagnósticos de Covid-19", type = "o",
 pch = 16)
>
```



Interpretação: pelo gráfico podemos ver que desde o começo do mês houve um aumento progressivo no número de casos diagnosticados, tendo como pico de casos entre os dias 15 a 20 do mês. Apesar das medidas serem tomadas no dia 15, alguns dias seguintes ainda havia aumento nos casos, porém pouco depois percebeu-se uma diminuição considerável na redução de diagnósticos, mostrando eficácia das medidas de controle.

Observação: Perceba que nessa questão não houve utilização de algum banco de dados do pacote "aplore3", devido ausência de variáveis relacionadas ao tempo, por isso, foi necessário a montagem de um banco de dados com esse tipo de variável para exemplificar o gráfico de linha. A construção do banco de dados foi feita no Excel, para saber mais detalhes de como utilizar bancos de dados do Excel, utilize a apostila "Análises em Bioestatística Básica: uma introdução ao software R".

8. Medidas de Variabilidade

Desvio padrão: mede a dispersão de cada valor do conjunto de dados em relação à média. É obtido pela raiz quadrada da variância.

Variância: é a quase média dos desvios padrão ao quadrado.

Coeficiente de variação: é uma medida adimensional calculada dividindo o desvio padrão pela média. Expressa o quanto da escala de medida (média) é ocupada pelo desvio-padrão.

Exercícios:

8.1) Dado o conjunto X = { 2, 5, 8, 9}, pede-se:

a) Calcule a média aritmética de X.

$$\overline{x} = \frac{2+5+8+9}{4} = 6$$

b) Calcule a variância amostral de X.

$$Var(X) = s^2 = \frac{(2-6)^2+(5-6)^2+(8-6)^2+(9-6)^2}{(4-1)} = 10$$

c) Quais elementos de X pertencem ao intervalo $[\bar{x} - s; \bar{x} + s]$?

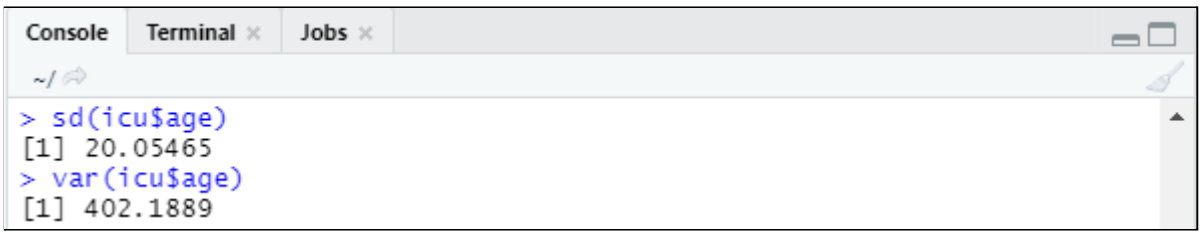
[6-($\sqrt{10}$); 6+ -($\sqrt{10}$)] = [2,84; 9,16]

Esse é o intervalo, portanto os elementos de X {2,5,8,9} que pertencem a esse intervalo são o 5 ; 8 e 9.

8.2) **(Banco de dados ICU - Pacote Aplore3)** Encontre uma medida que quantifique se existe grande variabilidade na idade dos pacientes.

Uma opção seria encontrar o desvio padrão ou variância, o que mostra se os dados possuem mais ou menos variabilidade entre eles.

Nesta questão, além da função "sd", usamos também a função "var" que é o comando para obter a variância. Isto é apresentado a seguir:



Interpretação: O desvio padrão da idade dos pacientes foi de 20 anos, isso significa que, em relação à média, as idades são de 20 anos menores ou maiores que a mesma; já a variância, que é o desvio padrão ao quadrado, é de 402.

8.3) **(Banco de dados ICU - Pacote Aplore3)** Supondo que um dos pesquisadores levantou a hipótese que as variáveis com menor variação estariam mais associadas à admissão do paciente na UTI, descubra entre as variáveis idade (age) e frequência cardíaca na admissão na UTI (hra), qual estaria mais associada à admissão do paciente na UTI segundo a hipótese do pesquisador.

Por se tratar de variáveis com unidades de medida diferentes precisamos utilizar o coeficiente de variação para obter uma medida de variabilidade em termos de percentual e assim poder compará-las.

Por meio de "sd", comando para obter o desvio padrão e "mean" para obter a média, fazemos a seguinte equação e dessa forma, obtemos o coeficiente de variação. Veja os comando utilizados:

Interpretação: aplicando a fórmula de coeficiente de variação, obtemos uma medida de variabilidade comparável entre as variáveis assim podemos dizer que a variável idade (age) tem uma maior variação do que a variável frequência cardíaca na admissão na UTI (hra), ou seja, a frequência cardíaca na admissão na UTI tem maior relação com a admissão do paciente na UTI, levando em consideração a hipótese dos pesquisadores.

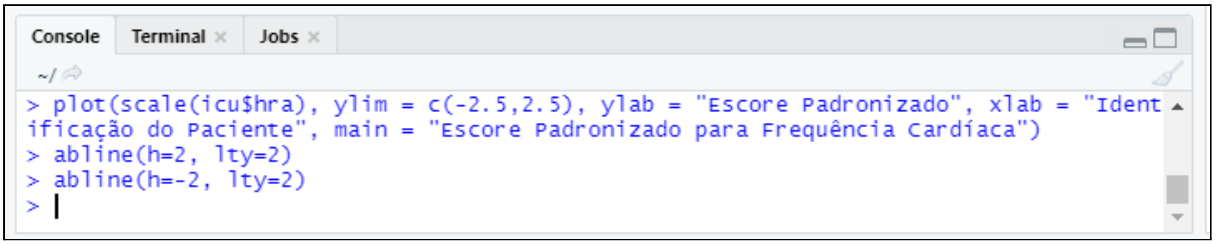
9. Identificando valores atípicos e Associação entre Variáveis

9.1) **(Banco de dados ICU - Pacote Aplore3)** A fim de monitorar os pacientes com maior alteração na frequência cardíaca, suponha que você é o responsável em identificar os pacientes com valores atípicos de frequência cardíaca. Como esses pacientes pondem ser identificados?

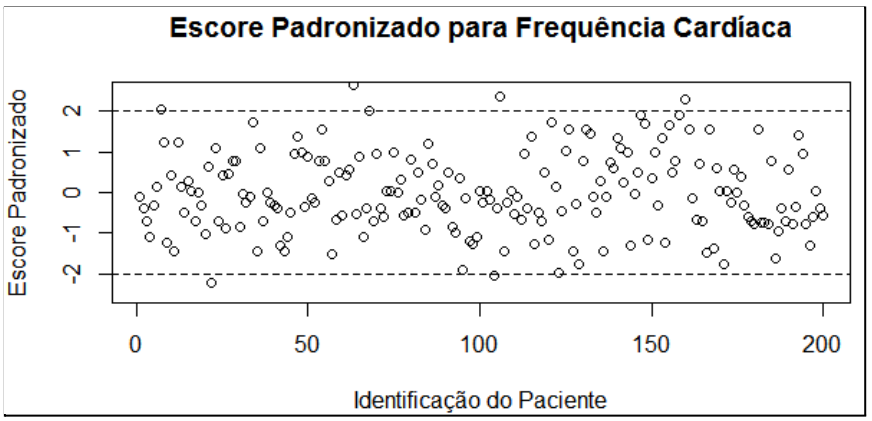
Poderíamos utilizar o escore padronizado para a variável frequência cardíaca na admissão na UTI, o que nos informa o número de desvios padrões que um valor dista da média. Abaixo apresentamos o escore padronizado para os 11 primeiros pacientes. O argumento "scale" é utilizado para obter o escore padronizado para cada valor da variável. No R temos então a seguinte imagem:

Interpretação: Considerando que valores usuais devem estar entre -2 e 2, o sétimo valor observado com escore de 2,05 é um valor que apresenta grande distância para a média amostral, ou seja, é um valor atípico. Por ser um valor positivo, isso indica que sua frequência cardíaca estava mais alta do que o normal.

Outra forma de identificar e apresentar esses valores de forma visual em gráfico, seria utilizando os seguintes comandos:



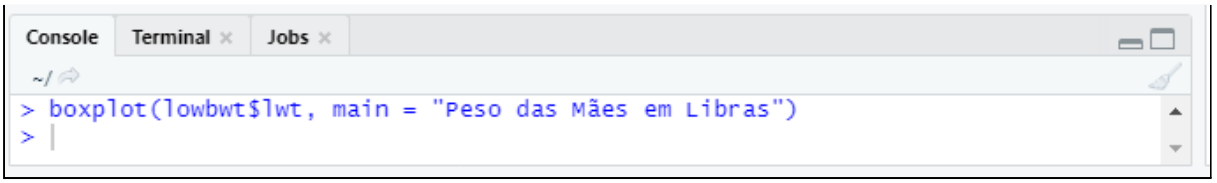
O Comando "plot(scale(icu$hra)" faz um gráfico de dispersão com base no escore padronizado; o argumento "ylim = c(-2.5,2.5)" determinará o tamanho do eixo y do gráfico; e tanto o "abline(h=2, lty=2)" quanto o "abline(h=-2, lty=2)" servem para adicionar as linhas horizontais, sendo que "lty" designa que será uma linha pontilhada.

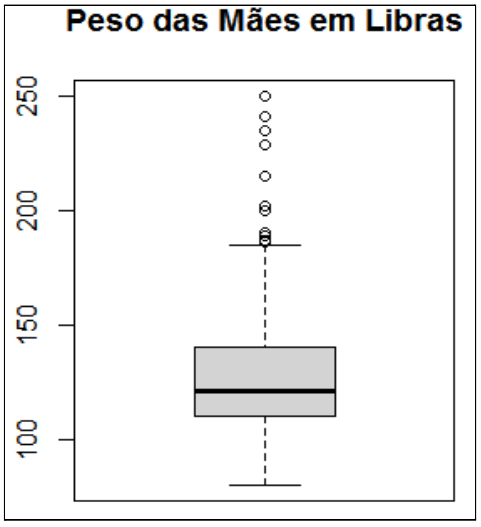


Interpretação: No gráfico é possível observar que 7 mães ultrapassaram a margem das linhas que representam os valores de corte para a determinação de um valor ser ou não considerado como atípico, sendo assim, concluímos que 7 pacientes possuíam frequência cardíaca atípica, quando comparada ao restante da amostra.

9.2) **(Banco de dados lowbwt - Pacote Aplore3)** Supondo que um nutricionista quer aplicar um programa de educação alimentar para as mães com maior peso dentre as participantes da pesquisa, e que a quantidade de vagas para o programa é bastante limitada. Encontre uma forma de identificar as mães com peso mais atípico em comparação às demais, para assim selecioná-las para o programa.

Uma forma de identificar esse grupo de mães é utilizando um boxplot, através do seguinte comando:





Interpretação: Pelo gráfico de boxplot, podemos verificar que as mães que se encontram com valores atípicos mais elevados são as que possuem aproximadamente peso acima de 180 libras.

9.3) **(Banco de dados ICU - Pacote Aplore3)** Supondo que na mesma população de estudo a idade e a frequência cardíaca em pacientes em leitos de internação possuem relação forte linear, o que preocupa os profissionais de saúde, descubra se o mesmo acontece nos paciente em UTI nesta amostra de pesquisa.

Podemos descobrir isso através do coeficiente de correlação de Pearson, verificando se existe uma relação linear entre as variáveis idade (age) e frequência cardíaca na admissão na UTI (hra).

Considere como parâmetro para classificar: 0<|r|<0,4 correlação fraca; 0,4=<|r|<0,7 correção moderada; 0,7=<|r|<1 correlação forte; e |r|=1 correlação perfeita.

No R a função usada para calcular a correlação entre as variáveis é a função "cor" com argumentos sendo as variáveis desejadas, como apresentamos abaixo:

Interpretação: Por ter apresentado um valor muito próximo de zero, a correlação entre essas duas variáveis é fraca, ou seja, não há evidências que exista relação linear entre as duas variáveis, diferente do ocorre com os pacientes em leitos de internação neste estudo.

10. Exercício de revisão

O conjunto de dados fornecido apresenta características de 19 pacientes com cisto no pâncreas:

Tabela de pacientes com cistos no pâncreas

Paciente	Sexo	Idade	Tamanho do cisto (cm)	Localização do cisto no pâncreas
1	F	49	6	CABEÇA
2	F	61	10	CABEÇA
3	M	34	8,2	CAUDA
4	F	73	3	COLO
5	M	47	3,6	CABEÇA
6	M	58	10	COLO
7	M	43	1	CABEÇA
8	M	71	1	CABEÇA
9	M	32	7	CAUDA

10	M	56	1	CABEÇA
11	M	61	6,6	CORPO
12	F	49	4	CABEÇA
13	M	80	3,1	CAUDA
14	M	72	2,3	CABEÇA
15	M	47	10,5	CABEÇA
16	F	48	6,5	CORPO
17	F	37	13	CORPO
18	M	71	1	COLO
19	M	74	7	CABEÇA

1) Construa uma tabela de frequências para a localização do cisto no pâncreas dos pacientes. Faça uma interpretação dos dados.

Tabela de frequência para região do cisto no pâncreas

LOCALIZAÇÃO	*FREQUÊNCIA*
Cabeça	*10 (52,9%)*
Corpo	*3 (15,7%)*
Colo	*3 (15,7%)*
Cauda	*3 (15,7%)*
TOTAL	*19 (100%)*

Dentre as possíveis localizações de cistos no pâncreas (cabeça, cauda, corpo e colo), a que mais é frequente entre os pacientes é na cabeça do pâncreas. Em relação às outras localizações, elas apresentam a mesma frequência (15,7%) entre esse grupo.

2) Construa uma tabela de frequências para o tamanho do cisto no pâncreas e o sexo dos pacientes;

Tabela de Frequência: Tamanho do Cisto vs Sexo dos Pacientes

TAMANHO DO CISTO (cm)	SEXO		TOTAL
	FEMININO	MASCULINO	
1 |- 5	2 (10,52%)	7 (36,84%)	9 (100%)
5 |- 10	3 (15,78%)	4 (21,05%)	7 (100%)
10 |- 15	1 (5,26%)	2 (10,52%)	3 (100%)
TOTAL	6 (31,57%)	13 (68,42%)	19 (100%)

a) Qual é a média de tamanho de cistos para os homens? E para as mulheres?

Cálculo para homens:

$$\overline{x} = \frac{8{,}2+3{,}6+10+1+1+7+1+6{,}6+3{,}1+2{,}3+10{,}5+1+7}{13} = 4,79.$$

Cálculo para mulheres:

$$\overline{x} = \frac{6+10+3+4+6{,}5+13}{6} = 7,08.$$

b) Qual é a mediana de tamanho de cistos para os homens? E para as mulheres?

Cálculo para homens:

Ordenando os valores em ordem crescente temos:

(1; 1; 1; 1; 2,3; 3,1; 3,6; 6,6; 7; 7; 8,2; 10; 10,5)

Para encontrar a mediana é só ordenar os valores de forma crescente, e pegar o valor central pois temos um total de 13 participantes homens. Assim, a mediana é igual a 3,6cm.

Cálculo para mulheres:

Ordenando os valores em ordem crescente temos:

(3 ;4; 6; 6,5; 10; 13)

Para encontrar a mediana é só ordenar os valores de forma crescente, verificar se o total de participantes mulheres é um número par ou ímpar. Como é um número par (6) basta fazer a média entre o 3º e 4º valores para obter a mediana. Ou seja, a mediana é igual a 6,25cm.

3) Construa uma tabela e um gráfico que permitam visualizar se há uma relação entre as variáveis tamanho do cisto no pâncreas e idade dos pacientes.
Obs: Para facilitar, crie uma tabela no excel e utilize-a no R. A explicação de como importar dados do excel para o R está detalhadamente feita na apostila "Análises em Bioestatística Básica: uma introdução ao software R"

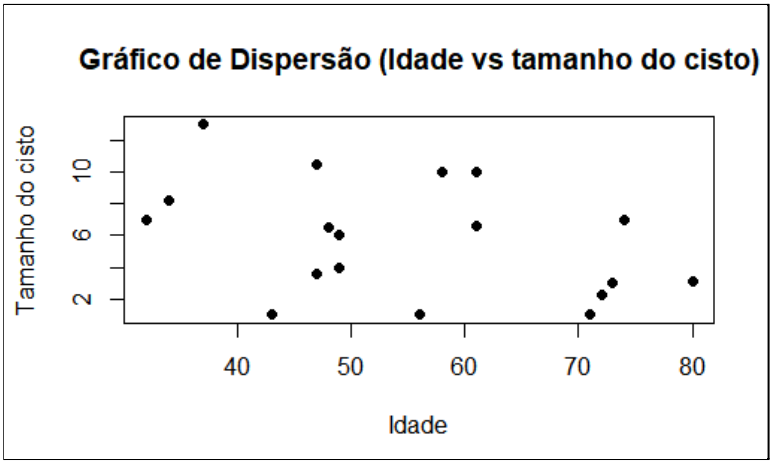
Tabela Tamanho dos Cistos x Idade dos Pacientes

TAMANHO DO CISTO (cm)	IDADE					TOTAL
	30-40	41-50	51-60	61-70	71-80	
1 |- 5	-	3 (15,78%)	1 (5,26%)	-	5 (26,31%)	9 (100%)
5 |- 10	2 (10,52%)	2 (10,52%)	-	1 (5,26%)	1 (5,26%)	6 (100%)
10 |- 15	1 (5,26%)	1 (5,26%)	1 (5,26%)	1 (5,26%)	-	4 (100%)
TOTAL	3 (15,78%)	6 (31,57%)	2 (10,52%)	2 (10,52%)	6 (31,57%)	19 (100%)

Para realizarmos esta questão pelo software R utilizaremos novamente a função "plot" (já explicada anteriormente).

```
Console  Terminal  Jobs
~/
> plot(cisto$Idade, cisto$`Tamanho do cisto (cm)`, main = "Gráfico de
 dispersão(idade vs tamanho do cisto)", ylab="Tamanho do cisto", xlab
="Idade", pch=19)
```

Obtemos o gráfico abaixo:



Interpretação: Não é possível observar nenhum padrão de comportamento indicando alguma relação entre as variáveis. Na prática isso significa que, pacientes mais velhos não necessariamente serão acometidos por um cisto maior.

4) Qual o desvio padrão do tamanho de cistos para esses pacientes?

No software R, utilizando o comando "sd" para obter o desvio padrão.

```
Console  Terminal  Jobs
~/
> sd(cisto$`Tamanho do cisto (cm)`)
[1] 3.690116
```

Interpretação: O desvio padrão do tamanho dos cistos para esses pacientes é bem grande, demonstrando que há bastante variação do tamanho destes cistos entre esses pacientes.

5) Anteriormente, foram fornecidas tabelas dos tamanhos dos cistos no pâncreas segundo o sexo e idade dos pacientes. Interprete- os.

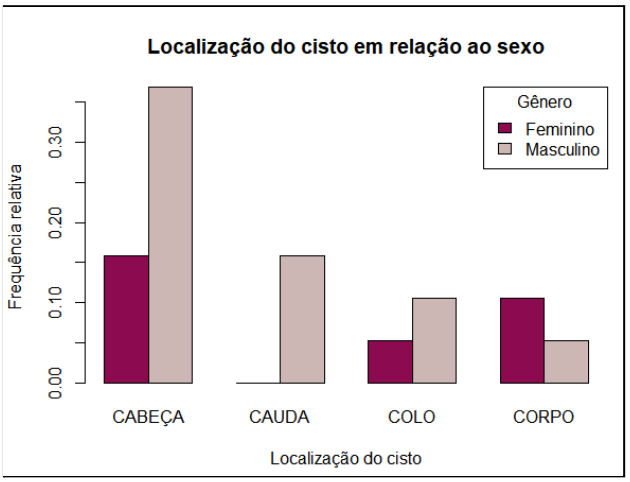
Entre os homens, a maior frequência de cistos no pâncreas são de tamanhos pequenos (de 1 a 5 cm), sendo mais frequente em homens mais velhos (71 a 80 anos). Já para as

mulheres, os cistos tendem a ser um pouco maiores (5 a 10 cm) e aparecem mais precocemente (30 a 50 anos).

6) Construa um gráfico que permita avaliar se existe relação entre o sexo do paciente e a localização do cisto. Comente o resultado.

Para investigar se existe alguma relação entre as variáveis podemos fazer um gráfico de barras com a função "barplot", visto que as duas variáveis são qualitativas. Para isso, vamos precisar usar o argumento "col" para definir as cores do gráfico, "legend" para criar a caixa de legenda e seus respectivos nomes, em que "topright" indica que a caixa de legenda ficará no canto superior direito e "fill" indica a cor de preenchimento.

```
> barplot(prop.table(table(cisto$Sexo,cisto$`Localização do cisto no pâncreas`)),
+ main="Localização do cisto em relação ao sexo",
+ beside= T, xlab= "Localização do cisto", col= c("deeppink4","mistyrose3"),
+ ylab= "Frequência relativa")
> legend("topright", legend= c("Feminino","Masculino"),
+ fill=c("deeppink4","mistyrose3"),title="Gênero")
```

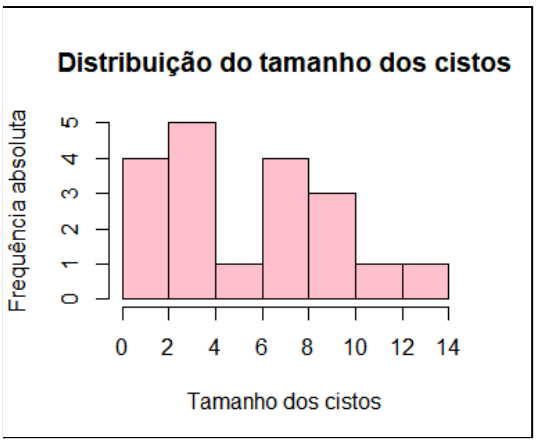


Interpretação: Notamos que, aparentemente, as duas variáveis estão relacionadas, visto que a distribuição de frequências da localização tem comportamento diferente para os dois sexos. Nas mulheres deste estudo, por exemplo, não foi observado cisto na cauda do pâncreas, enquanto que para ambos os sexos, a maior ocorrência do cisto é na cabeça do pâncreas.

7) Construa um gráfico de frequência absoluta do tamanho dos cistos no pâncreas.

Como trata-se de uma variável contínua, vamos construir um histograma utilizando a função "hist" do software R.

```
> hist(cisto$`Tamanho do cisto (cm)`, main="Distribuição do tamanho do
s cistos", xlab= "Tamanho dos cistos", ylab="Frequência absoluta", col
="pink")
```



Interpretação: Com base nos dados desses pacientes, podemos observar que a maior prevalência de tamanho de cistos são os de 0 a 4 cm. Há também uma prevalência significativa de cistos maiores, com tamanho de 6 a 10 cm, mas cistos muito grandes, ou seja, maiores que 10 não são tão frequentes.

11. Probabilidade

Alguns Conceitos:

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

Evento: é um subconjunto do espaço amostral relacionado a um experimento aleatório.

Exercícios:

11.1) Numa pesquisa feita com n = 20 crianças de uma escola foi encontrado que, dentre as três frutas da estação (banana, laranja e abacaxi), 10 delas têm como fruta favorita o abacaxi, 2 a laranja e 8 a banana. Para uma criança selecionada ao acaso, encontre a probabilidade do abacaxi ser a fruta favorita dessa criança.

Inicialmente, precisamos verificar o espaço amostral e o nosso evento de interesse. Como falado no enunciado, o espaço amostral consiste em todos os possíveis resultados, ou seja a preferência de cada aluno em relação às frutas da estação (banana, laranja e abacaxi). O evento, por sua vez, é a fruta abacaxi ser a favorita dentre as 3 possíveis.

A probabilidade será calculada através da divisão entre o número de vezes que o evento de interesse ocorre na amostra pelo número total de elementos na amostra. De tal forma, faremos a seguinte divisão:

Temos então que: $P(\text{abacaxi}) = \frac{10}{20} = 0.5 = 50\%$.

Interpretação: A probabilidade de se perguntar para uma dessas crianças qual a sua fruta favorita e ter como resposta "abacaxi" é de 50%.

11.2) **(Questão fictícia)** No contexto de pandemia do SARS-CoV-2, muito tem sido feito em busca de uma medida de prevenção ou diminuição dos casos clínicos dessa doença. Nesse sentido, um estudo feito com 13.060 voluntários, buscou averiguar a eficácia de uma vacina nacional. O delineamento do grupo de participantes desse estudo está representado na tabela abaixo. As colunas, da esquerda para a direita, representam o início do estudo, onde os critérios de inclusão e exclusão foram realizados; em seguida, o número de participantes que foram infectados; o número de infectados que desenvolveram sintomas leves e, por fim, o número de infectados com sintomas moderados a graves.

Delineamento dos voluntários do estudo

Grupo	Recrutados (total)	Infectados	Infectados com Sintomas leves	Infectados com Sintomas moderados a graves
Placebo	6530	167	138	37
Intervenção	6530	87	68	0
Total	13060	254	206	37

Pede-se:

a) Qual a probabilidade de um voluntário deste estudo ser infectado?

Definindo o evento ser infectado como I temos que:

$P(I)=\frac{254}{13060}= 0,0195$

Ou seja, a probabilidade de um voluntário deste estudo ser infectado é de 1,95%

b) Qual a probabilidade de desenvolver sintomas graves a moderados sabendo que é do grupo placebo?

Definindo o evento desenvolver sintomas graves como SG e grupo placebo como PL, temos que:

$$P(SG|PL) = \frac{P(SG \cap PL)}{P(PL)} = \frac{\frac{37}{13060}}{\frac{6530}{13060}} = \frac{37}{6530} = 0,0056$$

Ou seja, a probabilidade de desenvolver sintomas graves sabendo-se que a pessoa é do grupo placebo é de 0,56%

c) Sabendo-se que o voluntário é do grupo intervenção, qual a probabilidade de desenvolver sintomas graves?

Definindo o evento desenvolver sintomas graves como SG e grupo intervenção como IN, temos que:

$$P(SG|IN) = \frac{P(SG \cap IN)}{P(IN)} = \frac{\frac{0}{13060}}{\frac{6530}{13060}} = \frac{0}{13060} \times \frac{13060}{6530} = 0$$

Ou seja, se o voluntário é do grupo intervenção, a probabilidade de ele desenvolver sintomas graves a moderados é 0, ou seja, há 100% de probabilidade de seu quadro clínico não se desenvolver ao ponto grave.

d) Interprete os dados, tomando como base os seus conhecimentos de probabilidade.

Dentre os voluntários deste estudo, os que tomaram a vacina, ou seja, os do grupo intervenção, tiveram uma probabilidade menor de desenvolverem consequências de nível médio a moderado. Dessa forma, pode-se dizer, que pelo menos para este estudo, a intervenção foi benéfica pensando no desfecho clínico.

11.3) Uma amostra foi coletada em uma determinada cidade, e os indivíduos foram classificados com base em seu índice de massa corporal (IMC), além de terem sido separados por sexo. Os dados foram apresentados na tabela a seguir:

Estado nutricional dos indivíduos com base no IMC

Sexo	Baixo peso	Eutrófico	Sobrepeso	Obesidade	Total
Masculino	2	286	121	93	502
Feminino	17	294	140	74	525
Total	19	580	261	167	1027

Considere os seguintes eventos: M= Sexo masculino, F= Sexo feminino, B= Baixo peso, E= Eutrófico, S= Sobrepeso e O = Obesidade.

Se uma pessoa é sorteada aleatoriamente, calcule a probabilidade da pessoa sorteada:

a) Ser do sexo masculino e estar eutrófico

$P(M \cap E) = \frac{286}{1027} = 0{,}278.$

b) Estar com obesidade

$P(O) = \frac{167}{1027} = 0{,}163.$

c) Estar com sobrepeso ou ser do sexo masculino

$P(S \cup M) = P(S) + P(M) - P(S \cap M) = \frac{261}{1027} + \frac{502}{1027} - \frac{121}{1027} = 0{,}625.$

d) Estar com baixo peso, sabendo-se que é do sexo feminino

$P(B|F) = \frac{P(B \cap F)}{P(F)} = \frac{\frac{17}{1027}}{\frac{525}{1027}} = \frac{17}{525} = 0{,}032.$

12. Teorema de Bayes

Alguns Conceitos:

Probabilidade total: caso B_1, B_2,....B_n, seja uma partição do espaço amostral Ω e seja A um evento qualquer em Ω, a probabilidade de acontecer o evento A será o somatório das probabilidades de acontecer o evento A dado que aconteceu o evento B_i vezes a probabilidade de acontecer o evento B_i, veja:

$$P(A) = \sum_{i=1}^{k} P(A|B_i)\, P(B_i)$$

Teorema de Bayes: é um teorema para encontrar uma probabilidade condicional que utiliza o teorema da probabilidade total e a definição de probabilidade condicional. Seja B1, B2,... Bn uma partição do espaço amostral Ω e seja A um evento qualquer em Ω temos que:

$$P(B|A) = \frac{P(A|B)\, P(B)}{P(A)};\ \text{onde}\ P(A) = \sum_{i=1}^{k} P(A|B_i)\, P(B_i)$$

12.1) Para um dado tratamento medicamentoso, a probabilidade do indivíduo sofrer efeitos colaterais é de 0,70. Em um hospital, uma equipe médica consegue ter eficácia de 40% nesse tratamento quando o paciente apresenta efeito colateral e de 85% quando o paciente não apresenta efeitos colaterais. Considerando um paciente aleatório calcule: (considerando EC= efeito colateral e EF=eficácia do tratamento)

a) Qual a probabilidade da equipe médica ter eficácia no tratamento deste paciente?

$P(EF) = P(EF \cap EC) + P(EF \cap EC^c) = 0,4x0,7 + 0,85x0,3 = 0,28+0,255 = 0,535$

A probabilidade da equipe médica ter eficácia nesse tratamento medicamentoso, nesse paciente aleatório, sem antes saber se ele terá efeitos colaterais ou não, é de 53,5%.

b) Se o tratamento tiver sido eficaz, qual a probabilidade do paciente não ter apresentado efeitos colaterais?

$$P(EC^c|EF) = \frac{P(EC^C \cap EF)}{P(EF)} = \frac{0,85 \times 0,3}{0,535} = 0,477$$

Sabendo que o tratamento foi eficaz, a probabilidade do paciente não ter tido efeitos colaterais é de 47,7%.

12.2) Durante a pandemia da COVID-19, após a elaboração de vacinas eficientes, surgiu uma nova preocupação em relação à saúde pública e o processo de imunização em massa: a produção de seringas e agulhas. Existem 2 fábricas responsáveis pela produção de seringas, a fábrica A produz 60% da demanda, das quais 1% dessas seringas são defeituosas. Já a fábrica B produz os 40% restantes da demanda, sendo que 2% das seringas são defeituosas. (Considere para essa questão D representando o evento seringas defeituosas, e A e B para representando os eventos onde a seringa foi produzida pelas fábricas A e B respectivamente.

a) Qual a probabilidade de se ter uma seringa defeituosa?

$P(D) = P(D \cap A) + P(D \cap B)$

$P(D) = P(A)\,P(D|A) + P(B)\,P(D|B) = 0,01x0,60 + 0,02x0,40 = 0,014$

A probabilidade de ter uma seringa defeituosa em qualquer uma das fábricas é de 1,4%.

b) Considerando que a seringa é defeituosa, qual a probabilidade de ter sido produzida pela fábrica A?

$$P(A|D) = \frac{P(A \cap D)}{P(D)} = \frac{P(A)P(D|A)}{P(A)P(D|A)+P(B)P(D|B)} = \frac{0,01 \times 0,60}{0,01 \times 0,60 + 0,02 \times 0,40} = 0,429$$

Quando se tem uma seringa defeituosa, a probabilidade dela ter sido feita na fábrica A é de 42,9%.

c) Já no caso das agulhas, elas são produzidas por outras 3 fábricas, F, G e H. Na fábrica F se produz 40% das agulhas, na G 35% e na H 25%. A percentagem de agulhas defeituosas é de 1% para as fábricas F e G e de 1,5% para a fábrica H. Qual a probabilidade de selecionar uma agulha aleatoriamente e ela ser defeituosa? (Considere D=agulhas defeituosas, e F,G e H para representar a probabilidade das fábricas)

$P(D) = P(D \cap F)+P(D \cap G)+P(D \cap H)$

P(D)= P(D)P(D|F)+P(D)P(D|G)+P(D)P(D|H)= 0,01x0,40+0,01x0,35+0,015x0,25 = 0,0113

A probabilidade de se obter uma agulha defeituosa em alguma das fábricas é de 1,13%.

d) Sabendo que é uma agulha defeituosa qual a probabilidade de ter sido fabricada na fábrica G ou H?

P((G|D)∪(H|D))=

$$\frac{P(G)P(DG)+P(H)P(D|H)}{P(D)P(D|F)+P(D)P(D|G)+P(D)P(D|H)} = \frac{0{,}01\times0{,}35+0{,}015\times0{,}25}{0{,}01\times0{,}40+0{,}01\times0{,}35+0{,}015\times0{,}25} = 0,644$$

Sabendo que a agulha é defeituosa, a probabilidade dela ter sido feita na fábrica G ou na H é de 64,4%.

12.3) Determinado pássaro em um observatório pode ter problemas respiratórios ou físicos. Se ele tiver problemas respiratórios, é levado para o atendimento 24h, mas se tiver problema físico permanece no observatório. A chance desse pássaro ter problemas respiratórios é de 0,2. Já a chance do mesmo pássaro ter problemas físicos é de 0,15 se não houve problema respiratório precedente, e de 0,25 se houve problema respiratório precedente. Agora calcule:

a) Qual a probabilidade de o pássaro permanecer no observatório determinado dia?

Considere R sendo o evento ter problema respiratório e F sendo o evento ter problema físico, são dadas as seguintes informações:

P(R)= 0,2

$P(F|R^C)$= 0,15

P(F|R)= 0,25

A probabilidade do pássaro permanecer no observatório é a probabilidade de ele ter tido um problema físico, independente de ter havido um problema respiratório ou não.

P(F)= P(R) x P(F|R) + $P(R^C)$ x $P(F|R^C)$

P(F) = 0,2 x 0,25 + 0,8 x 0,15 = 0,17 ou 17%

b) Se o pássaro permaneceu no observatório em certo dia, qual a chance que tenha havido problema respiratório?

Devemos calcular a probabilidade de ter havido problema respiratório condicionada ao fato de sabermos que o pássaro permaneceu no observatório (lembrando que o pássaro permanece no observatório quando tem problema físico). Isso é feito por meio do Teorema de Bayes, veja:

$$P(R|F) = \frac{P(F|R)\,P(R)}{P(F)} = \frac{0{,}25 \times 0{,}2}{0{,}17} = 0,294 = 29,4\%$$

c) Qual a probabilidade de que tenha havido problema respiratório em determinado dia se o pássaro foi levado para o atendimento 24h?

Mais uma vez utilizamos o Teorema de Bayes para calcular a probabilidade de que tenha havido problema respiratório dado que o pássaro foi levado para o atendimento 24h. Consideramos NF o evento "não ter problema físico", temos que:

$$P(R|NF) = \frac{P(NF|R)\,P(R)}{P(NF)}$$

A probabilidade de não haver problema físico é calculada pela propriedade do evento complementar:

P(NF)= 1- P(F) = 1- 0,17 = 0,83

Como dito no enunciado, a probabilidade de ter um problema físico dado que houve um problema respiratório é de 0,25. A chance de não ter um problema físico será calculada por meio da propriedade do evento complementar, ou seja:

P(NF|R)= 1- P(F|R)= 1-0,25= 0,75

Substituindo na fórmula do teorema de Bayes temos:

$$P(R|NF) = \frac{0{,}75 \times 0{,}2}{0{,}83} = 0,181 = 18,1\%$$

13. Teste Diagnóstico

Alguns Conceitos:

Para o conteúdo a seguir, considere os seguintes eventos:

T+ o número de testes com resultados positivos;

D o número total de doentes;

T^- o número de testes com resultados negativos;

ND o número total de não doentes/sadios.

Sensibilidade: é a probabilidade do teste dar positivo para uma pessoa que está doente, ou seja, detectar uma pessoa como doente, dado que o indivíduo realmente está doente.

$$s = P(T^+|D) = \frac{P(T^+ \cap D)}{P(D)};$$

Especificidade: é a probabilidade do teste dar negativo para quem não está doente, ou seja, não detectar uma pessoa como doente, dado que o indivíduo não está doente.

$$e = P(T^-|ND) = \frac{P(T^- \cap ND)}{P(ND)};$$

Valor de predição positiva (VPP): é a probabilidade do indivíduo estar doente dado que o resultado do teste foi positivo. Ou seja, é a probabilidade de acerto do diagnóstico positivo é dada por:

$$VPP = P(D|T^+) = \frac{P(D \cap T^+)}{P(T^+)};$$

Conhecendo a sensibilidade, especificidade e prevalência da doença na população podemos calcular VPP pela seguinte equação:

$$VPP = \frac{sp}{sp + (1-e)(1-p)}.$$

Valor de predição negativa (VPN): é a probabilidade do indivíduo estar sadio, dado que o resultado do teste foi negativo. Ou seja, é a probabilidade de acerto do diagnóstico negativo.

$$VPN = P(ND|T^-) = \frac{P(ND \cap T^-)}{P(T^-)};$$

Conhecendo a sensibilidade, especificidade e prevalência da doença na população podemos calcular VPN pela seguinte equação:

$$VPN = \frac{e\,(1-p)}{e\,(1-p) + (1-s)p};$$

Este resultado é oriundo da aplicação do teorema de Bayes e deve ser utilizado quando o estudo é planejado para estimar bem sensibilidade e especificidade, mas não pode ser utilizado para estimar a prevalência da doença.

Proporção de Falsos Positivos (PFP): É a probabilidade do paciente não estar doente dado que o teste teve resultado positivo.

$$PFP = P(ND|T^+) = \frac{P(ND \cap T^+)}{P(T^+)} \quad \text{ou} \quad PFP = 1 - VPP$$

Proporção de Falsos Negativos (PFN): É a probabilidade do paciente estar doente dado que o teste teve resultado negativo.

$$PFN = P(D|T^-) = \frac{P(D \cap T^-)}{P(T^-)} \quad \text{ou} \quad PFN = 1 - VPN$$

Exercícios:

13.1) Embora o IMC seja um índice amplamente utilizado para detectar sobrepeso e obesidade na população, um estudo realizado com 388 indivíduos foi realizado a fim de verificar a acurácia desse índice quanto a detecção de excesso de adiposidade (gordura corporal). Assim, primeiramente foi medido o IMC do paciente e depois foram realizados exames laboratoriais e bioquímicos para saber se o paciente tinha, de fato, excesso de adiposidade. Os resultados estão apresentados abaixo:

Tabela de resultados do IMC na detecção de excesso de adiposidade

Categorização segundo IMC	Excesso de adiposidade segundo parâmetros bioquímicos e laboratoriais		Total
	Doente	Não doente	
Obeso (positivo)	163	5	168
Eutrófico (negativo)	144	99	243
Total	307	104	411

a) Calcule a sensibilidade do IMC com base nos resultados deste estudo.

$$s = P(T^+|D) = \frac{P(T^+ \cap D)}{P(D)} = \frac{163}{307} = 0,53 = 53\%$$

b) Calcule a especificidade do IMC com base nos resultados deste estudo.

$$e = P(T^{-}|ND) = \frac{P(T^{-} \cap ND)}{P(ND)} = \frac{99}{104} = 0,95 = 95\%$$

c) Interprete esses resultados. Qual atenção se deve ter ao usar esse índice para identificar obesidade metabólica?

Com base nos resultados deste estudo, o IMC é um índice com baixa sensibilidade mas alta especificidade. Isso quer dizer que ele não detecta tão bem pessoas doentes, mas acerta quanto às pessoas saudáveis. Assim, subestima o número de obesos metabólicos considerando-nos como eutróficos, sendo que, na verdade, são doentes. Por isso, para pessoas detectadas pelo IMC como "eutróficas" é interessante que se avalie outros parâmetros para se ter certeza que aqueles indivíduos não são doentes.

13.2) Resultados de um novo teste utilizado para diagnóstico de dengue mostraram que 82 dos testes de 98 pacientes com dengue foram positivos, 12 entre 98 eram limítrofes e quatro eram negativos. Em contraste, entre 387 indivíduos sem dengue, 8 foram positivos ao mesmo teste, 14 foram limítrofes e 365 foram negativos.

a) Inicialmente os autores calcularam a sensibilidade e a especificidade excluindo os resultados limítrofes. O que eles encontraram?

Excluindo os resultados limítrofes, temos 82 com resultado positivo em um grupo de 86 pacientes doentes (82 com resultado positivo e 4 negativos). Assim, calculamos a sensibilidade por:

$$s = P(T^{+}|D) = \frac{P(T^{+} \cap D)}{P(D)} = \frac{82}{86} = 0,95 \; ou \; 95\%$$

Pelo mesmo raciocínio, temos 365 com resultado negativo em um grupo de 373 indivíduos sadios (365 com resultado negativo e 8 com resultado positivo). Assim, a especificidade é calculada por:

$$e = P(T^{-}|ND) = \frac{P(T^{-} \cap ND)}{P(ND)} = \frac{365}{373} = 0,9785 = 97,85\%$$

b) Em seguida, os autores assumiram outra decisão: consideraram os resultados limítrofes como positivos. Qual a sensibilidade e a especificidade para essa nova situação? O que acontece com esses valores?

Para essa situação, considerando resultados limítrofes como positivos, temos um grupo de 98 pacientes com dengue, sendo que 94 possuem resultados positivos (82 com resultado positivo e 12 com resultado limítrofe que agora é considerado positivo também). Assim, a sensibilidade é calculada por:

$s = P(T^+|D) = \frac{P(T^+ \cap D)}{P(D)} = \frac{94}{98} = 0,96 = 96\%$

Para a especificidade, por sua vez, temos 365 negativos em um grupo de 387 saudáveis (consideramos aqui todos os indivíduos sem dengue).

$e = P(T^-|ND) = \frac{P(T^- \cap ND)}{P(ND)} = \frac{365}{387} = 0,94 = 94\%$

Assim, com essa decisão, a sensibilidade do teste aumentou, enquanto a especificidade diminuiu.

c) Depois, os autores assumiram uma decisão oposta: consideraram os resultados limítrofes como negativos. Qual a sensibilidade e a especificidade para essa nova situação? O que acontece com esses valores?

Para essa decisão, considerando resultados limítrofes como negativos, temos um grupo de 98 pacientes com dengue, sendo que 82 possuem resultados positivos. Assim, a sensibilidade é calculada por:

$s = P(T^+|D) = \frac{P(T^+ \cap D)}{P(D)} = \frac{82}{98} = 0,84 = 84\%$

Para a especificidade, por sua vez, temos 379 negativos (365 negativos e 14 com resultado limítrofe que agora é considerado como negativo), em um grupo de 387 saudáveis.

$e = P(T^-|ND) = \frac{P(T^- \cap ND)}{P(ND)} = \frac{379}{387} = 0,9793 = 97,93\%$

Assim, com essa decisão, a sensibilidade do teste diminuiu, enquanto a especificidade aumentou ligeiramente.

d) Comente os efeitos sobre a sensibilidade e a especificidade do novo teste nas três decisões relatadas acima.

Os melhores resultados simultâneos encontrados para sensibilidade e especificidade ocorreram quando foram descartados os resultados limítrofes. Ao se considerar os resultados limítrofes como positivos, há uma perda na especificidade, porque aqueles resultados limítrofes que na verdade forem negativos, serão considerados como positivos, ou seja, serão falsos positivos. Finalmente, ao se considerar os resultados limítrofes como negativos, há uma perda de sensibilidade, pois aqueles resultados

limítrofes que na verdade forem positivos, serão considerados como negativos, ou seja, serão falsos negativos.

e) Qual dessas opções seria melhor utilizar em uma conduta médica sabendo que na região há muitos casos de dengue?

A melhor opção seria considerar os resultados limítrofes como positivos, porque foi nessa situação que a sensibilidade do teste foi maior. Assim, na prática, esse teste acusaria, com mais precisão, doentes como positivos. É verdade que algumas pessoas saudáveis poderiam ser detectadas com resultado positivo para o teste (falsos positivos), mas como na região os casos de dengue são frequentes, seria melhor superestimar doentes que subestimar e não tratá-los de forma apropriada.

13.3) Em um estudo sobre o teste ergométrico, Wriner et al. (1979) compararam os resultados obtidos entre 1023 indivíduos com e 442 sem doença coronariana. O teste foi definido como positivo se foi observado mais de 1mm de depressão ou elevação do segmento ST, por pelo menos 0,08s, em comparação com os resultados obtidos com o paciente em repouso. O diagnóstico definitivo (classificação como doente ou não-doente) foi feito através de angiografia (teste padrão ouro).

Resultados obtidos pelo teste ergométrico com pacientes com e sem doença coronariana

Doença Coronariana	Teste ergométrico		
	T+	T-	Total
D+	815	208	1023
D-	115	327	442
Total	930	535	1465

a) Qual a especificidade do teste?

$$e = P(T^- | ND) = \frac{P(T^- \cap ND)}{P(ND)} = \frac{327}{442} = 0,74 = 74\%$$

b) Qual a sensibilidade do teste?

$$s = P(T^+ | D) = \frac{P(T^+ \cap D)}{P(D)} = \frac{815}{1023} = 0,79 = 79\%$$

13.4) Um exame físico foi usado para rastrear câncer de próstata em 2.500 homens com comprovação de biópsia da próstata e em 5.000 homens não doentes chamados de grupo controle, cuja raça e idade correspondiam ao par homens intervenção (aqueles que passaram pelo exame físico). Os resultados do exame físico de toque foi positivo em 1.800

casos e em 800 homens controle, todos os quais não mostraram evidência de câncer na biópsia.

a) Nesse experimento, qual foi a sensibilidade do exame físico?

$s = P(T^+|D) = \frac{P(T^+ \cap D)}{P(D)} = \frac{1800}{2500} = 0,72 = 72\%$

b) E a especificidade?

$e = P(T^-|ND) = \frac{P(T^- \cap ND)}{P(ND)} = \frac{4200}{5000} = 0,84 = 84\%$

c) Qual foi o valor preditivo positivo encontrado? Sabe-se que a prevalência da doença na população é de 11%.

$VPP = \frac{sp}{sp+(1-e)(1-p)} = \frac{0,72 \times 0,11}{0,72 \times 0,11+(1-0,84)(1-0,11)} = \frac{0,0792}{0,0792+(0,16)(0,89)} = \frac{0,0792}{0,0792+0,1424} = 0,02216 = 22,16\%$

d) Calcule o valor preditivo positivo tendo como base a prevalência da doença encontrada entre os participantes desse estudo.

$p = \frac{2500}{7500} = 0,33 = 33\%$

$VPP = \frac{sp}{sp+(1-e)(1-p)} = \frac{0,72 \times 0,33}{0,72 \times 0,33+(1-0,84)(1-0,33)} = \frac{0,2376}{0,2376+(0,16)(0,67)} = \frac{0,2376}{0,2376+0,1072} = 0,6891 = 68,91\%$

e) Por que os valores encontrados na alternativa c e d são tão destoantes?

Os valores de predição positiva foram tão destoantes, porque inicialmente a prevalência da doença foi dada no enunciado, e, por isso, o VPP refletiu melhor o resultado na população. Já na questão d, usamos a prevalência da doença com base nos valores fixados de doentes e sadios da tabela deste estudo, e assim, fizemos um recorte pequeno cujo resultado não condiz com a prevalência real da doença na população.

13.5) (VUNESP) A eficácia de um teste de laboratório para checar certa doença nas pessoas que comprovadamente têm essa doença é de 90%. Esse mesmo teste, porém, produz um falso positivo da ordem de 1%. Em um grupo populacional em que a incidência dessa doença é de 0,5%, seleciona-se uma pessoa ao acaso para fazer o teste. Qual a probabilidade de que o resultado desse teste venha a ser positivo?

s =P(T+|D+)=0,9 P(T+|D-)=0,01 P(T-|D+)=0,1 e= P(T-|D-)=0,99

Com base no enunciado, temos que a sensibilidade é de s =P(T+|D+)=0,9 =90%, a proporção de falsos positivos é de P(T+|D-)=0,01= 1% e que a prevalência de P(D+)=0,5%.

Portanto, a proporção de falsos negativos é de P(T-|D+)=10%, e a especificidade é de e= P(T-|D-)=99%. Sabendo que a proporção de doentes na população é de 0,5%, a proporção de indivíduos sadios é de 99,5%.

Assim, para calcularmos a probabilidade do resultado ser positivo, independente do indivíduo estar doente ou sadio, precisamos multiplicar a sensibilidade pela prevalência da doença nesse grupo e em seguida somar com a proporção de falsos positivos multiplicada pela proporção de indivíduos sadios na população, pois

$P(T+) = P(T+ \cap D+)$*+P(T+∩D-) = P(T+|D+)P(D+) + P(T+|D-)P(D-) = 0,90 × 0,005 + 0,01 × 0,995 = 0,01445 = 1,445%.*

13.6) Um questionário avaliou o consumo diário de vegetais auto referido pelos estudantes de uma escola e referido por seus pais em 500 alunos. Posteriormente foi realizado um recordatório de 24 horas em 3 dias diferentes e foi observado que 300 indivíduos realmente consumiam vegetais diariamente. Os resultados das avaliações estão representados nas tabelas abaixo:

Tabela de resultados auto referidos

Auto referido	R24 positivo	R24 negativo
Consome diariamente	240	40
Não consome diariamente	60	160

Tabela de resultados referidos pelos pais

Referido pelos pais	R24 positivo	R24 negativo
Consome diariamente	270	60
Não consome diariamente	30	140

Obs: O Recordatório Alimentar é uma ferramenta muito utilizada por nutricionistas para coleta de dados sobre a alimentação do paciente e avaliação do conteúdo calórico e de nutrientes que geralmente é consumido por esse indivíduo.

a) Calcule a sensibilidade para os dois tipos de avaliação.

Inicialmente iremos calcular a sensibilidade da avaliação auto referida, sendo C^+ o resultado auto relatado de consumo diário positivo e P o resultado positivo do teste R24h, temos que:

$$s = P(C^+|P) = \frac{P(C^+ \cap P)}{P(P)} = \frac{240}{300} = 0,80 \; ou \; 80\%$$

Agora para a sensibilidade da avaliação referida pelos pais, temos que:

$$s = P(C^+|P) = \frac{P(C^+ \cap P)}{P(P)} = \frac{270}{300} = 0,90 \; ou \; 90\%$$

b) Calcule a especificidade para os dois tipos de avaliação.

Inicialmente iremos calcular a especificidade da avaliação auto referida, sendo C^- o resultado auto relatado de não consumo diário e N o resultado negativo do teste R24h, temos que:

$$e = P(C^-|ND) = \frac{P(C^- \cap N)}{P(N)} = \frac{160}{200} = 0,80 \; ou \; 80\%$$

Agora para a especificidade da avaliação referida pelos pais, temos que:

$$e = P(C^-|ND) = \frac{P(C^- \cap N)}{P(N)} = \frac{140}{200} = 0,70 \; ou \; 70\%$$

c) Qual a proporção de falsos positivos dessas avaliações?

Inicialmente iremos calcular a proporção de falsos positivos da avaliação auto referida, sendo C^+ o resultado auto relatado de consumo diário e N o resultado negativo do teste R24h, temos que:

$$PFP = P(N|C+) = \frac{P(N \cap C^+)}{P(C^+)} = \frac{40}{280} = 0,14 \; ou \; 14\%$$

Agora para a proporção de falsos positivos da avaliação referida pelos pais, temos que:

$$PFP = P(N|C+) = \frac{P(N \cap C^+)}{P(C^+)} = \frac{60}{330} = 0,18 \; ou \; 18\%$$

Veja que a proporção de falsos positivos para a avaliação dos pais é maior que a auto avaliação das crianças, revelando que há uma tendência maior de relatos verdadeiros para o não consumo de vegetais pelas crianças que pelos seus responsáveis.

d) Qual a proporção de falsos negativos dessas avaliações?

Inicialmente iremos calcular a proporção de falsos negativos da avaliação auto referida, sendo C^- o resultado auto relatado de não consumo diário e RP o resultado positivo do teste R24h, temos que:

$$PFN = P(RP|C^-) = \frac{P(RP \cap C^-)}{P(C^-)} = \frac{60}{220} = 0,27 \; ou \; 27\%$$

Agora para a proporção de falsos negativos da avaliação referida pelos pais, temos que:

$$PFN = P(RP|C^-) = \frac{P(RP \cap C^-)}{P(C^-)} = \frac{30}{170} = 0,18 \; ou \; 18\%$$

Veja que a proporção de falsos negativos para a autoavaliação das crianças é maior que a avaliação dos seus pais, revelando que há uma tendência maior de relatos verdadeiros para o consumo de vegetais pelos seus pais do que pelas crianças.

14. Medidas de Associação

Alguns Conceitos:

Incidência: corresponde aos novos casos/indivíduos que surgem a partir de um determinado período. Exemplo: Indivíduos que foram diagnosticados com COVID-19 hoje.

Prevalência: se refere a todos os casos/indivíduos que já apresentam a ocorrência do evento a ser estudado/analisado, ou seja, a soma de todos os casos em um determinado período de tempo. Exemplo: Todos os indivíduos que foram diagnosticados com COVID-19 em Minas Gerais desde o início da pandemia até o fim de 2020.

As medidas de associação são muito utilizadas em estudos epidemiológicos na área de saúde. Dentre elas, as aqui apresentadas utilizam razão entre os números encontrados na pesquisa, normalmente resumidos em uma tabela 2x2, para estabelecer associação entre os resultados. Veja uma tabela genérica:

Tabela 2 x 2 genérica		
Grupos	Ocorrência do evento (Desfecho)	Total

	Sim	Não	
Grupo Exposto	a	b	a+b
Grupo Não Exposto	c	d	c+d
Total	a+c	b+d	a+b+c+d

Risco Relativo (RR): é a razão da probabilidade estimada de ocorrência do evento de interesse no grupo de expostos com esta probabilidade no grupo de não expostos. Essa medida é usada para calcular se um dos grupos tem um risco maior ou menor com base em sua exposição e seu desfecho. O RR só deve ser usado em estudos observacionais prospectivos ou experimentais (se estes forem aleatorizados), ou seja, que partiram da exposição e houve um acompanhamento até o desfecho, portanto avalia a incidência. O risco relativo pode ser calculado por:

$$RR= \frac{\left(\frac{a}{a+b}\right)}{\left(\frac{c}{c+d}\right)}.$$

Para a interpretação do RR vamos considerar que RR>1 significa que a incidência/risco no grupo exposto é maior em relação ao não exposto; RR=1 mostra que a incidência entre os grupos é igual; RR<1 significa que a incidência/risco no grupo exposto é menor em relação ao não exposto (quando se trata de uma doença indica fator de proteção).

Razão de Prevalência (RP): nessa medida, como o próprio nome diz, será uma conta realizada para saber a razão de prevalência do desfecho entre grupos exposto e não exposto. Perceba que a fórmula para se encontrar a RP é a mesma da de RR, no entanto, o nome é mudado pois nesta se trata de prevalência e não da incidência de novos casos, na prática, elas são usadas em tipos diferentes de estudo, sendo que a RP pode ser usada, por exemplo, em estudos do tipo observacional transversal. Veja a fórmula a seguir:

$$RP= \frac{\left(\frac{a}{a+b}\right)}{\left(\frac{c}{c+d}\right)}.$$

Na interpretação da RP vamos considerar que RP>1 significa que a prevalência da doença no grupo exposto é maior em relação ao grupo não exposto; RP=1 mostra que a prevalência entre os grupos é igual; RP<1 indica que a prevalência no grupo exposto é menor em relação ao grupo não exposto.

Razão de Chances(RC) / Odds Ratio(OR): em estatística chance é a probabilidade de um evento ocorrer dividido pela probabilidade desse mesmo evento não ocorrer. A RC ou OR, é usada para calcular se a chance de ocorrer o evento é maior, menor ou igual no grupo exposto em comparação ao grupo não exposto. Muito usada em estudos

caso-controle, mas também pode ser utilizada em outros estudos. Além disso, em algumas situações o RC pode também ser uma boa estimativa do RR, para isso os dados precisam ter as seguintes condições: a frequência da ocorrência nos casos/doença seja menor que 10%, os casos com a ocorrência do do evento sejam representativos de todos os casos com o evento ocorridos na população de onde eles foram retiradas, e os controles também representativos do grupo de pessoas que não ocorreu o evento na população de onde os casos foram retirados, ou seja, as amostras devem ser representativas da população. Portanto temos que a fórmula para OR é:

$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}.$$

Para interpretar o resultado temos que: OR>1 a exposição aumenta a chance para a ocorrência do evento; OR=1 não há relação entre exposição e desfecho; OR<1 a exposição diminui a chance para a ocorrência do evento.

Exercícios:

14.1) Supondo que em uma pesquisa de coorte, pesquisadores avaliaram a incidência de Diabetes Mellitus tipo 2, comparando pessoas que bebiam bebidas açucaradas com alta frequência de consumo (3 ou mais vezes na semana), e pessoas com baixa frequência de consumo (1 vez na semana ou menos de 4 vezes no mês). Pessoas com consumo intermediário não foram incluídas na comparação. Obtiveram a seguinte tabela:

Tabela de consumo de bebidas açucaradas e incidência de de Diabetes Mellitus tipo 2

Frequência de consumo	Desenvolvimento de Diabetes Mellitus tipo 2		Total
	Sim	Não	
Alta	461	15394	15855
Baixa	132	15761	15893
Total	593	31155	31748

a) Qual a medida mais apropriada para fazer uma associação dos dados?

A medida mais apropriada seria a de risco relativo, por se tratar de um estudo de coorte e partir da incidência, o que possibilita atribuir risco ou não à exposição.

b) Utilizando a medida de associação escolhida na questão anterior, o que podemos dizer sobre o desenvolvimento de Diabetes Mellitus tipo 2 e a frequência de consumo de bebidas açucaradas ?

Utilizando a fórmula de RR temos que :

$$RR = \frac{\left(\frac{461}{461+15394}\right)}{\left(\frac{132}{132+15761}\right)} = \frac{0{,}0291}{0{,}0083} = 3{,}51$$

Isso significa que o risco de desenvolver Diabetes Mellitus tipo 2 é 3,51 vezes maior no grupo de pessoas que consomem bebidas açucaradas com alta frequência, em comparação aos que consomem com baixa frequência.

14.2) Sabendo que o consumo de sal, consequentemente de sódio, entre os brasileiros é bastante acima do recomendado, e que o sódio está associado a hipertensão, um ensaio clínico controlado, avaliou se a substituição do sal refinado branco, pelo sal light (onde existe uma substituição de 50% de sódio por potássio) influenciaria na incidência de hipertensão. Os indivíduos inicialmente não apresentavam hipertensão, tinham de 50 a 60 anos e foram separados de forma aleatória em dois grupos, um manteve o consumo de sal refinado branco, e o outro foi orientado a utilizar apenas o sal light, os grupos foram acompanhados ao longo de 10 anos. Pacientes com doença renal foram excluídos do estudo devido a restrição do consumo de potássio. Os dados foram inseridos na tabela abaixo:

Tabela de incidência de hipertensão com base no tipo de sal utilizado

Tipo de sal consumido	Desenvolvimento de hipertensão		Total
	Sim	Não	
Sal light	83	2045	2128
Sal refinado branco	169	2163	2332
Total	252	4208	4460

a) Nesse estudo qual seria a medida mais apropriada para realizar uma associação entre os resultados?

Como temos um estudo que parte da exposição e avalia a incidência, o mais interessante é utilizar o risco relativo.

b) A partir do cálculo da medida de associação, o que podemos dizer com relação ao consumo de sal light nesse estudo?

Por meio da fórmula de RR obtemos o seguinte resultado:

$$RR = \frac{\left(\frac{83}{83+2045}\right)}{\left(\frac{169}{169+2163}\right)} = \frac{0{,}0390}{0{,}0725} = 0,54$$

Como o RR foi menor que 1, nesse caso, isso significa que o uso de sal light tem 46% de redução no risco de hipertensão, quando comparado ao sal refinado branco.

14.3) Um estudo com o objetivo de conhecer o perfil de pessoas com osteoporose, e se havia diferença entre os sexos, realizou um estudo observacional transversal, utilizando um banco de dados terciário para estudar a prevalência de osteoporose em homens e mulheres idosos (maior ou igual a 65 anos). Após terem recebidos os dados, organizaram a seguinte tabela:

Tabela de prevalência de osteoporose

Sexo biológico	Presença de osteoporose		Total
	Sim	Não	
Feminino	58	1845	1903
Masculino	31	1852	1883
Total	89	3697	3786

a) Qual medida de associação você usaria para avaliar esses dados?

A melhor medida de associação a ser utilizada nessa pesquisa seria a razão de prevalência, pois é a que melhor se aplica ao modelo de pesquisa, por se tratar de dados onde os indivíduos já "sofreram" a exposição, que neste caso, seria uma característica biológica, e em seguida observa-se qual foi a prevalência da doença. Outra opção seria a RC, mas como o estudo quer saber a prevalência, essa medida não vem ao caso.

b) Utilizando a medida de associação escolhida na questão anterior, o podemos aferir desses dados?

Através da fórmula de RP temos que:

$$RP = \frac{\left(\frac{58}{58+1845}\right)}{\left(\frac{31}{31+1852}\right)} = \frac{0{,}0305}{0{,}0165} = 1,848$$

Esse resultado diz que a prevalência de osteoporose em mulheres idosas é de 1,85 vezes maior do que em homens idosos.

14.4) Um grupo de pesquisa realizou um estudo transversal observacional, buscando saber se a prevalência de sobrepeso e obesidade diferenciava com base no ambiente alimentar dos indivíduos. Esse grupo tinha acesso às informações sobre o estado nutricional e a localização dos indivíduos de uma determinada cidade, a hipótese do estudo era de que os indivíduos que tivessem maior acesso a feiras livres e sacolões, logo de alimentos mais saudáveis, teriam uma menor prevalência de sobrepeso e obesidade. Por terem acesso a localização, puderam conferir a proximidade entre a residência dos indivíduos com os estabelecimentos de comida, separando assim, os participantes em dois grupos, com alto acesso e baixo acesso a alimentos saudáveis. Com os dados coletados, montou-se a seguinte tabela:

Tabela de prevalência de sobrepeso e obesidade

Acesso a alimentos saudáveis	Em estado de sobrepeso ou obesidade		Total
	Sim	Não	
Alto	25	147	172
Baixo	32	138	170
Total	57	285	342

a) Qual medida de associação deve-se utilizar nesta pesquisa?

Deve-se usar a razão de prevalência, como o próprio enunciado diz, a pesquisa quer avaliar justamente a prevalência, além disso, por se tratar de um estudo observacional transversal justifica ainda mais o uso dessa medida.

b) Após calcular a associação entre as variáveis, o que pode-se dizer sobre os dados da pesquisa?

Usando a fórmula de RP obtemos:

$$RP = \frac{\left(\frac{25}{25+147}\right)}{\left(\frac{32}{32+138}\right)} = \frac{0{,}1453}{0{,}1882} = 0{,}7721$$

Por termos obtido uma RP menor que 1, isso nos diz que a prevalência de pessoas com sobrepeso e obesidade no grupo de pessoas com alto acesso a alimentos saudáveis foi 23% menor quando comparado a pessoas com baixo acesso a esses alimentos.

14.5) Em busca de verificar se o consumo frequente de álcool estava relacionado com câncer de estômago, foi realizada uma pesquisa com pacientes de uma determinada cidade que apresentavam esse tipo de câncer e também pacientes que não apresentavam. Os indivíduos foram separados em dois grupos após terem realizado uma anamnese sobre o consumo de álcool. Obteve-se então a seguinte tabela:

Tabela presença de câncer x Consumo frequente de álcool

Consome álcool frequentemente	Possui câncer de estômago		Total
	Sim	Não	
Sim	59	417	476
Não	24	464	488
Total	83	881	964

a) Com base nas informações, qual seria a medida de associação mais apropriada?

Nesse caso, seria mais interessante usar OR, visto que o estudo busca saber se existe relação entre a frequência do consumo de álcool e o fato de possuir câncer de estômago.

b) Utilizando a medida de associação, o que pode-se dizer com relação ao consumo de álcool frequente?

Usando a fórmula de OR temos o seguinte resultado:

$$OR = \frac{\frac{59}{417}}{\frac{24}{464}} = \frac{59 \times 464}{417 \times 24} = 2,74$$

O valor encontrado significa que beber álcool com frequência leva a uma chance de 2,74 vezes maior de se desenvolver câncer de estômago, quando comparado a pessoas que não consomem frequentemente.

c) Com base nesses dados, pode-se dizer que a RC é uma boa estimativa do RR?

Não podemos afirmar isso, pois apesar do número de pessoas com câncer ser menor que 10% da amostra total, o enunciado não informa se os dados dessa amostra são representativos da população.

14.6) Médicos veterinários que estudam comportamento de animais domésticos (de pequeno porte), a fim de descobrir situações que contribuem para a depressão desses animais, realizaram um estudo caso-controle com cães, onde avaliaram a presença da companhia de outro cachorro e estado de depressão, separando os animais em dois grupos: viviam sem a presença de mais um cão, viviam com a presença de outro cão. Após a coleta de dados, formulou-se a seguinte tabela:

Tabela de cães com depressão

Convive com outro cão	Possui depressão		Total
	Sim	Não	
Sim	12	67	79
Não	18	62	80
Total	30	129	159

a) Com base nos dados e no enunciado, indique uma medida de associação interessante para se aplicar a esse estudo?

Uma boa medida seria a RC, pois estabeleceria uma relação entre o convívio com outro cão e a presença de depressão, sendo o que a pesquisa busca saber, também pelo fato de ser uma pesquisa de caso-controle.

b) Realize o cálculo da medida escolhida e interprete o resultado.

Aplicando a fórmula temos que:

$$OR = \frac{\frac{12}{67}}{\frac{18}{62}} = \frac{12 \times 62}{67 \times 18} = 0,62$$

Com base na medida de associação, podemos dizer que neste estudo, encontrou-se que cães que possuem convívio com outro cão têm uma chance 38% menor de se ter depressão do que cães que convivem sem outro cão.

15. Distribuições de Probabilidades ou Modelos Probabilísticos

A **distribuição de probabilidade ou modelo probabilístico** serve para descrever a probabilidade de ocorrência dos possíveis valores que uma variável aleatória pode assumir. Uma distribuição de probabilidade pode ser apresentada por meio de fórmulas, tabelas ou gráficos.

Alguns Conceitos:

Variável aleatória: é uma variável numérica, ou seja quantitativa (qualitativas podem ser transformadas) onde seu resultado será um valor aleatório. Essas variáveis podem ser classificadas em discretas ou contínuas. O mais comum é representá-las com letras maiúsculas.

I) Modelos para variáveis aleatórias discretas

Variável aleatória discreta: será uma variável discreta quando o valor for um número enumerável, sem casas decimais. Uma probabilidade Y, terá como possíveis valores y_1, y_2, y_3….

Modelo Uniforme Discreto: todos os possíveis valores ocorrem com a mesma probabilidade $P(X = x) = 1/k$, em que k é a quantidade de possíveis respostas.

Modelo Binomial: se baseia nos ensaios de Bernoulli, onde os experimentos aleatórios podem ter como resposta 0 e 1, fracasso e sucesso ou sim e não.

$$P(X = x) = p(x) = \frac{n!}{x!(n-x)!} \times p^{x} \times (1 - p)^{n-x},$$

em que pé a probabilidade de “sucessos”, χ é o número de sucessos em n ensaios e n é o número de ensaios.

Exercícios:

15.1) Dos experimentos a seguir, verifique quais são binomiais, identificando, quando possível, os valores dos parâmetros n e p. Para aqueles que não são binomiais, justifique.

a) De uma sala com cinco mulheres e três homens, selecionar, aleatoriamente e com reposição, três pessoas. A variável aleatória de interesse é o número de mulheres selecionadas na amostra.

É binominal porque cada ensaio tem apenas dois resultados, mulher e homem, os resultados de cada seleção são aleatórios e a probabilidade de retirar uma mulher é constante. Os parâmetros da distribuição são n = 8 e p = $\frac{5}{8}$.

b) De uma população de milhares de pessoas de um hospital, selecionar aleatoriamente e sem reposição, vinte pessoas. O interesse está no número de doentes na amostra.

Não é distribuição binomial porque os eventos não são independentes entre si, já que não há reposição dessas pessoas. No entanto, se o tamanho da população é muito grande se comparado a tamanho amostral, cálculos aproximados podem ser feitos assumindo independência e o modelo binomial.

c) Selecionar uma amostra aleatória simples de 500 pessoas no estado de Minas Gerais. O interesse está no número de favoráveis à mudança da capital do município de Belo Horizonte para cidades do interior do estado.

É binominal porque cada ensaio (resposta de cada pessoa) tem apenas dois resultados possíveis, favorável e desfavorável, e os resultados dos ensaios são independentes entre si. Os parâmetros são n = 500 e p é a proporção de respostas favoráveis na população.

d) Observar uma amostra aleatória simples de 100 crianças recém-nascidas no Hospital das Clínicas de Belo Horizonte. A variável aleatória em questão é o peso, em kg, de cada criança da amostra.

Não é uma distribuição binomial porque existem diversos resultados possíveis para o peso de cada criança (ex: 2.0kg, 2.1kg, 2.3kg, 4.0kg….).

15.2) Em um determinado hospital foi estipulado a distribuição de probabilidade de se receber pacientes com queimaduras durante uma semana:

Números de pacientes com queimadura (X)	0	1	2	3
$p_i = P(X = x_i)$	10/20	4/20	4/20	2/20

Um enfermeiro precisa conhecer melhor a probabilidade de se ter queimados no hospital para organizar as equipes e insumos hospitalares, para isso calcule as seguintes probabilidades:

a) A probabilidade de se ter pelo menos 1 paciente com queimadura no hospital em uma semana.

$P(X \geq 1) = P(X=1) + P(X=2) + P(X=3) = \frac{4}{20} + \frac{4}{20} + \frac{2}{20} = \frac{1}{2}$

A probabilidade de se ter pelo menos um paciente com queimadura no hospital em uma semana é de 0,5 ou 50%.

b) A probabilidade de se ter mais de dois pacientes com queimadura em uma semana.

$P(X \geq 2) = P(X=3) = \frac{2}{20} = \frac{1}{10}$

A probabilidade de se ter dois pacientes com queimaduras no hospital em uma semana é de 0,1 ou 10%.

c) O número esperado (médio) semanal de pacientes com queimaduras no hospital.

$E(X) = 0 \times P(X=0) + 1 \times P(X=1) + 2 \times P(X=2) + 3 \times P(X=3) = \left(0 \times \frac{10}{20}\right) + \left(1 \times \frac{4}{20}\right) + \left(2 \times \frac{4}{20}\right) + \left(3 \times \frac{2}{20}\right) = 0,9$

O número de pacientes esperado em uma semana é de 0,9.

15.3) Supondo que 2 a cada 5 pessoas com diabetes, possui alguma complicação progressiva da doença (nefropatia, neuropatia, doenças cardiovascular ou retinopatia), e que estes vão precisar de tratamento medicamento, caso 12 pacientes com diabetes em um hospital, descubra as seguintes probabilidades:

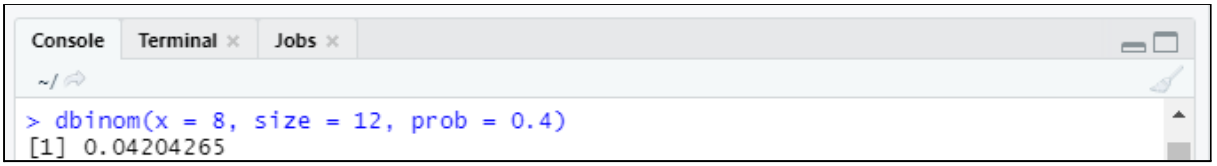
a) A probabilidade de que exatamente 8 tenha alguma complicação progressiva.

O número de pessoas com alguma complicação progressiva pode ser modelado com o modelo Binomial com parâmetros n = 12 e p = ⅖, e assim:

$P(X=8) = \frac{12!}{8!(12-8)!} \times \left(\frac{2}{5}\right)^8 \times \left(\frac{3}{5}\right)^4 = 0,042.$

A probabilidade de que exatamente 8 dos 12 pacientes com diabetes tenham alguma complicação progressiva da diabetes é de 0,042 ou 4,2%.

Realizando pelo R temos que:



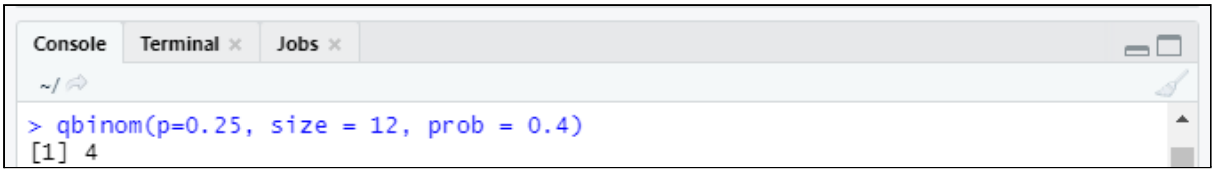
b) Considerando um grupo de 60 pessoas com diabetes, qual o número esperado de pessoas que desenvolveram complicações progressivas da doença:

$E(X) = 60 \times \frac{2}{5} = 24$

O valor esperado de pacientes com complicações progressivas da diabetes em um grupo de 60 diabéticos é de 24 indivíduos.

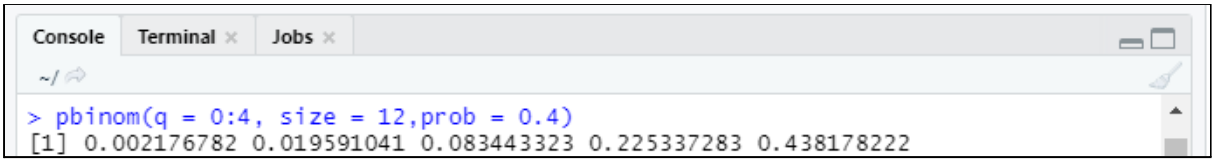
c) Através do R é possível também calcular com simplicidade um percentil da distribuição binomial. Encontre percentil de ordem 25% do número de doentes com complicações, isto é, o valor $P_{25\%}$ tal que a probabilidade de ter no máximo $P_{25\%}$ doentes com complicações é de pelo menos 25% e a probabilidade de ter no mínimo $P_{25\%}$ doentes com complicações é de pelo menos 75%.

Para resolver essa questão utilizamos os comandos abaixo, onde o termo "qbinom" é referente ao quantil/ordem de percentil da binomial e o "p=0.25" indicará o percentual. Observe:



Interpretação: Isso significa que com 25% de probabilidade teremos no máximo 4 sucessos, que neste caso, o "sucesso" representa o número de pessoas com diabetes que iram ter complicações progressivas da doença.

Caso não houvesse essa função, a forma de encontrar o número de sucessos de um percentil específico seria através do cálculo da probabilidade acumulada para que assim pudéssemos ver qual valor seria encontrado no percentil de ordem $P_{25\%}$ o que demandaria bastante tempo dependendo da situação. Através do R poderia ser calculado essa probabilidade acumulada através do comando "pbinom", veja:



O argumento "0:4" calcula todos os valores de uma só vez o que agiliza o processo, porém ainda assim não é prático como a função "qbinom", uma vez que precisamos avaliar em qual intervalo se encontra o valor do percentil de ordem $P_{25\%}$, que neste caso é o valor 4, visto que o valor 3 se encerra em $P_{23\%}$, e o valor 4 é encontrado até no percentil de ordem $P_{44\%}$, portanto englobando o percentil de ordem $P_{25\%}$.

15.4) Um certo programa de treinamento físico melhora o rendimento de 60% das pessoas a ele submetidas. Qual a probabilidade de, numa amostra de 8 pessoas que sejam submetidas a esse programa, menos da metade tenha seu rendimento melhorado?

P(X<4)= P(X=0) + P(X=1) + P(X=2) + P(X=3)

P(X=0) = $\frac{8!}{0!(8-0)!}$ *x (0,6)^0 x (0,4)^8 = 1 x 1 x 0,00065536 = 0,00065536*

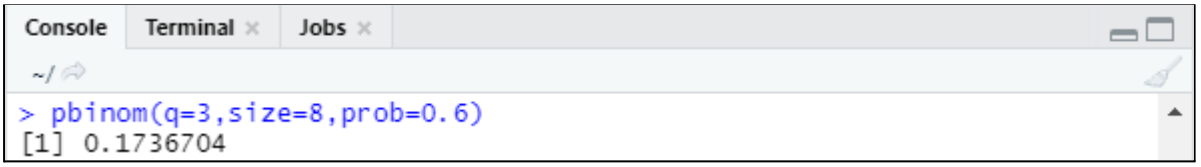
P(X=1) = $\frac{8!}{1!(8-1)!}$ *x (0,6)^1 x (0,4)^7 = 8 x 0,6 x 0,0016384= 0,00786432*

P(X=2) = $\frac{8!}{2!(8-2)!}$ *x (0,6)^2 x (0,4)^6 = 28 x 0,36 x 0,004096= 0,04128768*

P(X=3) = $\frac{8!}{3!(8-3)!}$ *x (0,6)^3 x (0,4)^5 = 56 x 0,216x 0,01024= 0,12386304*

P(X<4)= 0,00065536 + 0,00786432 + 0,04128768 + 0,12386304 = 0,1736704 ou 17,37%

No software R podemos obter a probabilidade pedida através do comando abaixo::

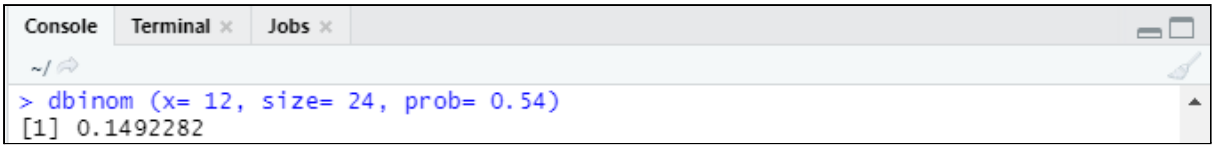


15.5) (Usando o R) Num estudo realizado com brasileiros foi observado que 54% dos jovens com idade inferior a 18 anos consumiam, ou consumiram, em algum momento, bebida alcoólica. Se escolhermos, aleatoriamente, 24 adolescentes dessa faixa etária, qual a probabilidade de:

a) 12 jovens já terem consumido bebida alcoólica?

Para realização no programa R a função "dbinom" calcula a probabilidade de uma variável aleatória com distribuição binomial assumir determinado valor x, "x" se refere ao número de jovens que consomem bebida alcoólica, para o qual queremos calcular uma

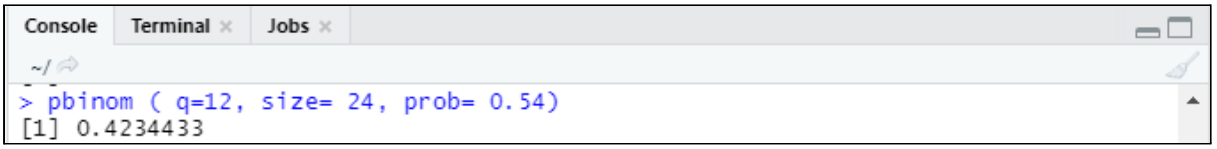
probabilidade, "size" é o tamanho da amostra e "prob" é a proporção de jovens brasileiros que consomem bebida alcoólica.



Ou seja, a probabilidade de 12 jovens, dentre esses escolhidos, já terem consumido bebida alcoólica é de 14,92%.

b) No máximo 12 jovens já terem consumido bebida alcoólica?

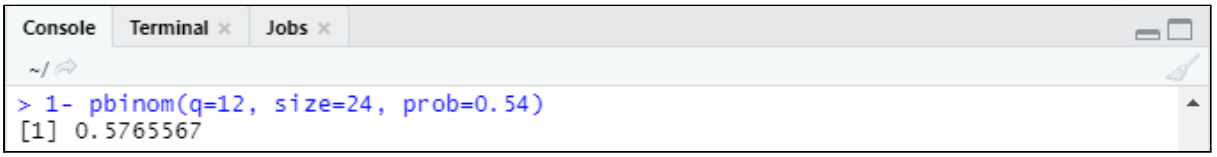
Já para essa questão utilize "pbinom" que é a função para calcular a probabilidade acumulada em uma distribuição binominal, "q" se refere ao número máximo jovens que queremos encontrar.



Ou seja, a probabilidade de até 12 jovens, dentre esses escolhidos, já terem consumido bebida alcoólica é de 42,34%.

c) Mais do que 12 jovens já terem consumido bebida alcóolica?

Para isso, utilizamos o comando "pbinom" que calcula a probabilidade acumulada até o valor q em uma distribuição binominal. A probabilidade pedida no enunciado é do evento complementar ter no máximo 12 jovens que consumiram bebida alcoólica.

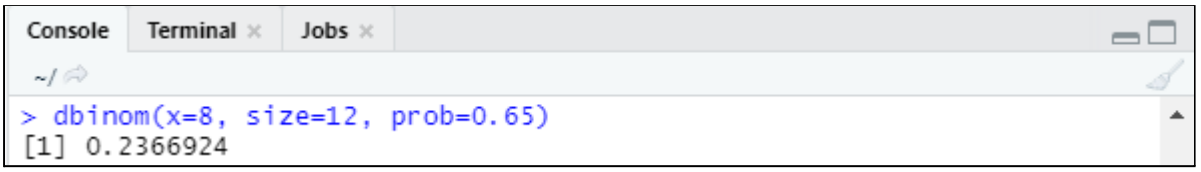


Ou seja, a probabilidade de mais de 12 jovens já terem consumido bebida alcoólica é de 57,65%.

15.6) (Usando o R) A probabilidade de um paciente de UTI de levar alta para o quarto em um período de 7 dias é de 0,65. Qual é a probabilidade de saírem:

a) Exatamente 8 pacientes dentre os atuais 12 pacientes dessa UTI num período de até 7 dias?

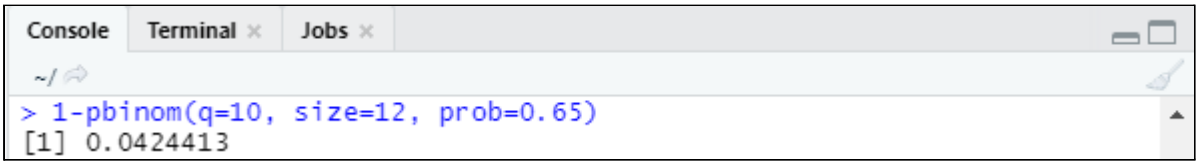
Para realização no programa R utilizamos o comando "dbinom" que já foi explicado anteriormente.



Ou seja, a probabilidade de que exatamente 8 pacientes, dentre os 12, saiam da UTI num período de até 7 dias é de 23,67%

b) Mais de 10 pacientes em um total de 12 pacientes dessa UTI num período de até 7 dias?

Como queremos saber a probabilidade de mais de 10 pacientes saírem da UTI, temos que subtrair da frequência acumulada de até 10 pacientes. Assim, utilizamos o comando "pbinom" cujos argumentos foram explicados anteriormente e subtraímos do total (1=100%).



Assim, a probabilidade de mais de 10 pacientes saírem da UTI num período máximo de até 7 dias é de 4,24%.

II) Modelos para variáveis aleatórias contínuas

Alguns Conceitos:

Variável aleatória contínua: os possíveis valores são não enumeráveis, consequentemente a probabilidade calculada sobre um único valor é igual a zero, sendo necessário calcular probabilidade em intervalos através da função positiva denominada densidade de probabilidade.

Modelo Normal ou Gaussiano: é a variável aleatória contínua cuja função de densidade é dada por:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$, para $-\infty \leq x \leq +\infty$,

em que μ e σ^2 são os parâmetros de média e variância, respectivamente.

Distribuição Normal Padrão: é a distribuição normal com média (μ) igual a zero e variância (σ^2) igual a 1.

Exercícios:

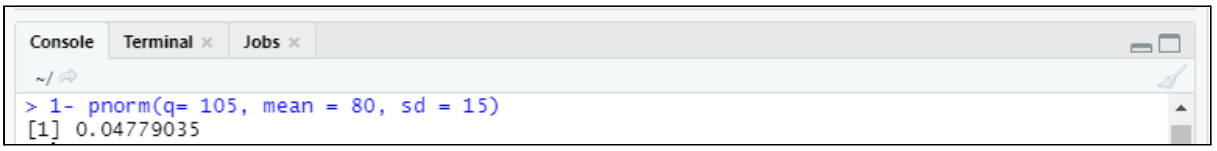
15.7) A distribuição do nível de glicemia em jejum dos idosos de um instituto de longa permanência é normal com média 80mg/dL e desvio padrão de 15 mg/dL.

a) Qual a probabilidade de um desses idosos, selecionado ao acaso, ter glicemia em jejum acima de 105 mg/dL?

$$P(X > 105) = P\left(\frac{X-\mu}{\sigma} > \frac{105-\mu}{\sigma}\right) = P\left(Z > \frac{105-80}{15}\right) = P(Z > 1{,}66) = 0{,}0485.$$

Ou seja, a probabilidade de um desses idosos ter glicemia acima de 105 mg/dL é de 4,85%

Para solucionar questões de distribuição normal por meio do software R, iremos utilizar o comando "pnorm", em que o argumento "q" indicará o valor utilizado no cálculo para encontrar o ponto da curva na distribuição normal, e o "sd" diz o desvio padrão. O "1 - " no começo do comando tem a mesma função usada nas questões de binomial, de calcular a probabilidade do evento complementar, mas também poderíamos usar o comando "lower.tail = FALSE" ao invés de "1 - ".



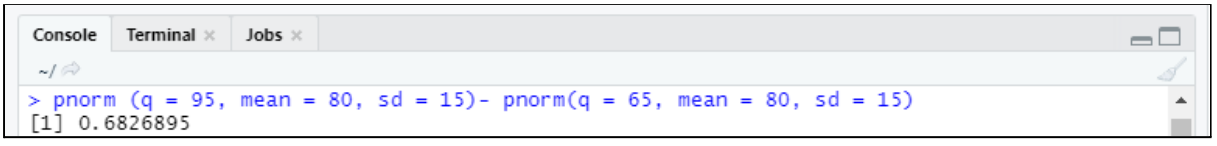
b) Qual a porcentagem esperada de idosos com glicemia entre 65 e 95 mg/dL?

$$P(65 \leq X \leq 95) = P\left(\frac{65-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{95-\mu}{\sigma}\right) = P\left(\frac{65-80}{15} \leq Z \leq \frac{95-80}{15}\right) =$$

$$= P(-1 \leq Z \leq +1) = 0{,}8413 - 0{,}1587 = 0{,}6826$$

Ou seja, a probabilidade de um desses idosos ter glicemia entre 65 e 95 mg/dL é de 68,26%

Realizando no R temos:



c) Qual o valor para o qual 90% dos idosos tenha glicemia inferior a ele?

Procuramos o valor de y tal que P (X < y) = 0,90.

Padronizando temos:

$$P\left(\frac{X-\mu}{\sigma} < \frac{y-\mu}{\sigma}\right) = P\left(Z < \frac{y-80}{15}\right) = 0{,}90.$$

Portanto, $\frac{y-80}{15}$ *= 1,28, logo y = (1,28 x 15) + 80 = 99,2 mg/dL.*

Ou seja, 90% dos idosos têm glicemia em jejum inferior a 99,2 mg/dL.

Para realização no R, utilizaremos a função qnorm para encontrar o percentil de ordem 90% da distribuição normal de forma semelhante a função "qbinom" nas questões de binomiais.

```
Console  Terminal  Jobs
~/
> qnorm (p = 0.9, mean = 80, sd = 15)
[1] 99.22327
```

15.8) Suponha que numa certa região, o peso dos homens adultos tenha distribuição normal com média 70 kg e desvio padrão 16 kg. E o peso das mulheres adultas tenha distribuição normal com média 60 kg e desvio padrão 12 kg. Ao selecionar uma pessoa ao acaso, o que é mais provável: uma mulher com mais de 75 kg ou um homem com mais de 90 kg?

Para homens:

$$P(X>90) = P\left(\frac{X-\mu}{\sigma} > \frac{90-\mu}{\sigma}\right) = P\left(Z > \frac{90-70}{16}\right) = P(Z > 1{,}25) = P(Z < -1{,}25) = 0{,}1056.$$

Para mulheres:

$$P(X>75) = P\left(\frac{X-\mu}{\sigma} > \frac{75-\mu}{\sigma}\right) = P\left(Z > \frac{75-60}{12}\right) = P(Z > 1{,}25) = P(Z < -1{,}25) = 0{,}1056.$$

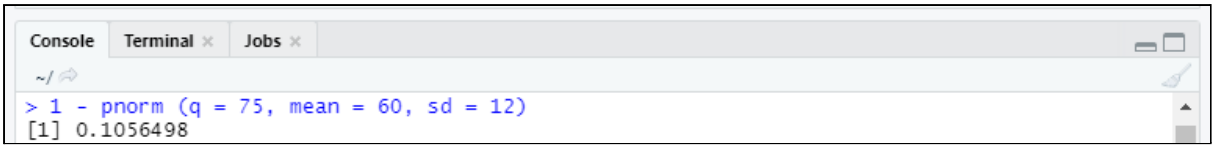
Isso significa que a probabilidade de se encontrar um homem com peso maior que 90kg é igual a probabilidade de se encontrar uma mulher com peso maior que 75kg.

Fazendo no R temos os seguintes comandos:

Para o peso dos homens:

```
Console  Terminal  Jobs
~/
> 1 - pnorm (q = 90, mean = 70, sd = 16)
[1] 0.1056498
```

Para o peso das mulheres:



15.9) Um teste padronizado é aplicado a um grande número de praticantes de atividade física. Os seus resultados são normalmente distribuídos com média de 50 pontos e desvio padrão de 10 pontos. Se Joana conseguir 65 pontos, qual é a percentagem esperada de praticantes de atividade física com mais pontos do que Joana?

Calculamos esta percentagem através da probabilidade de um resultado maior do que 65 pontos.

$P(X>65) = P\left(\frac{X-\mu}{\sigma} > \frac{65-\mu}{\sigma}\right) = P\left(Z > \frac{65-50}{10}\right) = P(Z> 1,5) = P(Z<-1,5) = 0,0668.$

Isso significa que a percentagem esperada de participantes que façam mais pontos que Joana é de 6,68%.

Solucionando essa questão no R temos:

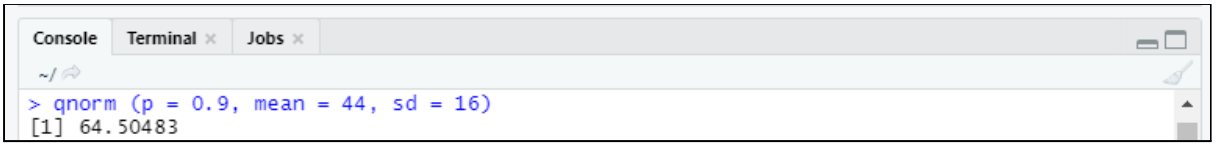
15.10) Calculou-se em 44 minutos o tempo médio que os funcionários de uma empresa levam para almoçar, com desvio padrão de 16 minutos. Quanto deve ser a duração do intervalo de almoço, de modo a permitir tempo suficiente para que 90% dos trabalhadores terminem o almoço em tempo inferior a ele? Admita distribuição normal para o tempo de duração do almoço.

Precisamos calcular y tal que $P(X < y) = 0,90$.

Assim, $P\left(\frac{X-\mu}{\sigma} < \frac{y-\mu}{\sigma}\right) = P\left(Z < \frac{y-44}{16}\right) = 0,90$. *Com isso obtemos* $\frac{y-100}{15} = -1,28$ *e portanto,* $y = (1,28 \times 16) + 44 = 64,48$ *minutos.*

Ou seja, o tempo necessário, em minutos, para que 90% dos funcionários terminem o almoço antes do fim, é de 64 minutos e 30 segundos.

A resolução no R dessa questão é da seguinte forma:



15.11) Na distribuição normal é possível construir faixas de referência, um dos métodos é pela curva de Gauss. Sabendo disso, para uma população onde se tem distribuição normal, a média de expectativa de vida é de 75 anos, com desvio padrão de 10 anos, solucione as seguintes questões:

a) Construa uma faixa de referência para 95% desses indivíduos.

$FR(\gamma) = (\mu - z_{\gamma}\sigma;\ \mu + z_{\gamma}\sigma)$

$P(Z < z_{\gamma}) = \gamma + \frac{(1-\gamma)}{2} = 0{,}95 + \frac{(1-0{,}95)}{2} = 0{,}975.$

Quando observamos o valor na tabela normal, vemos que 0,975 corresponde a 1,96, portanto esse será o valor utilizado para construção da faixa de referência.

Assim, obtemos

$FR(\gamma) = (75 - 1{,}96 \times 10;\ 75 + 1{,}96 \times 10) = (55{,}4;\ 94{,}6).$

Essa faixa significa que 95% dos indivíduos dessa população terão sua expectativa de vida entre os valores de 55,4 anos a 94,6 anos.

b) Para uma certa amostra de 9 indivíduos, qual a probabilidade de terem uma expectativa de vida entre 80 a 90 anos de idade ?

Note que $\bar{X} \sim N\left(\mu;\ \frac{\sigma^2}{n}\right) = \left(75;\ \frac{10^2}{9}\right)$. *Assim, temos que*

$P(80 < \bar{X} < 90) = P(\bar{X} < 90) - P(\bar{X} < 80)$

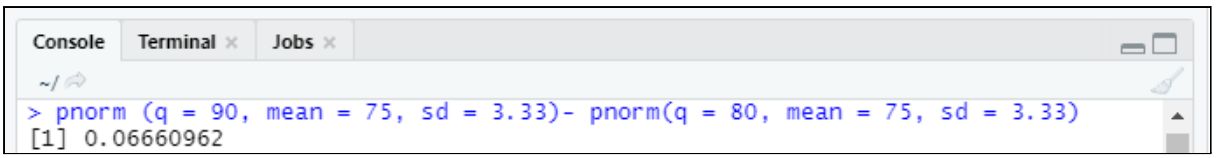
$P(80 < \bar{X} < 90) = P\left(Z < \frac{90-75}{3{,}33}\right) - P\left(Z < \frac{80-75}{3{,}33}\right)$

$P(80 < \bar{X} < 90) = P(Z < 4{,}5) - P(Z < 1{,}50)$

$P(80 < \bar{X} < 90) = 1 - 0{,}9332 = 0{,}0668$ ou 6,68%

Isso significa que a probabilidade de 9 pessoas aleatórias terem a expectativa de vida entre 80 a 90 anos é de 6,68%.

A resolução no R se dá pela seguinte maneira:



15.12) Quando já conhecemos os valores de uma variável de uma população, que tenha distribuição normal, podemos calcular a probabilidade da média amostral. Suponha que em uma cidade sua população tenha uma N(9; 6) para a variável casos de dengue por semana, retirando uma amostra de 15 indivíduos de um bairro dessa cidade, encontre a probabilidade de média amostral ser:

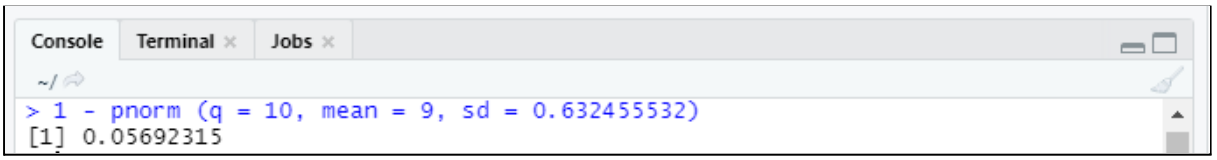
X~N(9; 6) n = 15 $\rightarrow$ $\overline{X}$~N(9; $\frac{6}{15}$)

a) Maior do que 10.

$$P(\overline{X}>10) = P\left(\frac{\overline{X}-9}{\sqrt{\frac{6}{15}}} > \frac{10-9}{\sqrt{\frac{6}{15}}}\right) = P\,(Z>1{,}581) = 1 - 0{,}9429 = 0{,}0571 \text{ ou } 5{,}71\%$$

A probabilidade da média amostral ser maior do que 10 casos de dengue é de 5,71%.

Sendo que no R o comando é:



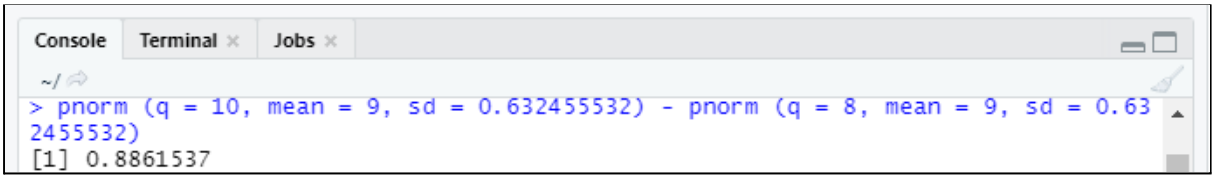
b) Estar entre 8 e 10.

$$P(8 < \overline{X} < 10) = P(\overline{X} < 10) - P(\overline{X} < 8) = P\left(Z < \frac{10-9}{\sqrt{\frac{6}{15}}}\right) - P\left(Z < \frac{8-9}{\sqrt{\frac{6}{15}}}\right)$$

$P(8 < \overline{X} < 10) = P(Z<1{,}581) - P(Z<-1{,}581) = 0{,}9429 - 0{,}0571 = 0{,}8858.$

A probabilidade da média amostral estar entre 8 e 10 casos de dengue é de 88,58%.

Resolução no R:

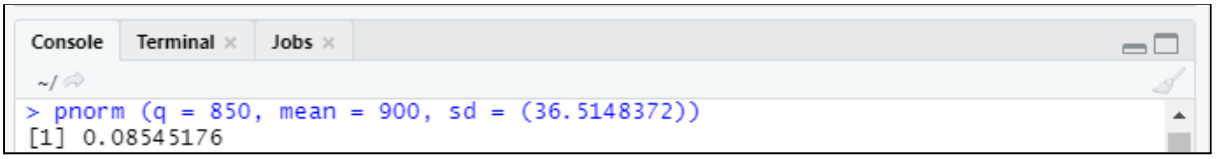


15.13) Uma empresa que cria ratos de laboratório, trabalha com um espécie que possui expectativa de vida média de 900 dias, com desvio padrão de 200 dias. Uma universidade comprou 30 ratos dessa empresa. Suponha que a afirmação da empresa seja verdadeira.

a) Qual a probabilidade de que a expectativa de vida dessa amostra seja menor que 850 dias ?

$$P(\bar{X}<850)= P\left(Z < \frac{850-900}{\frac{200}{\sqrt{30}}}\right) = P(Z < 1,369) = 1 - 0,9131 = 0,0869.$$

A probabilidade da expectativa de vida média para uma amostra de 30 ratos dessa espécie, ser menor que 850 dias, é de 8,69%.



b) O que foi necessário supor para solucionar a questão anterior?

Foi realizada supondo que a amostra foi suficientemente grande para aplicar o Teorema Central do Limite.

15.14) Suponha que uma pesquisa tenha encontrado que a suplementação de 50mg de ferro por um período de 3 meses, é suficiente para tratar a anemia ferropriva de 80% dos pacientes, sendo que os demais devem manter o tratamento ou utilizar uma dose maior. Considerando uma amostra de 30 indivíduos com anemia ferropriva, qual a probabilidade de menos de 70% destes indivíduos terem um tratamento eficaz? Considere que essa amostra seja suficientemente grande para aplicar o TCL.

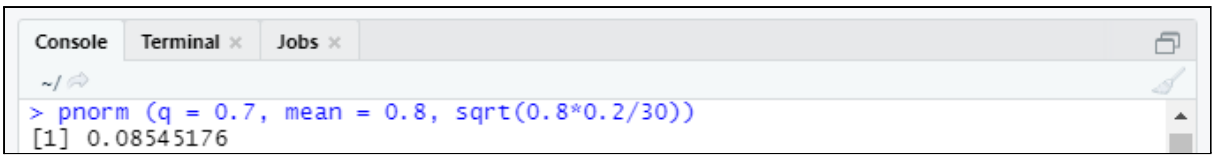
$n=30 \qquad \hat{p} \sim N\left(0,80; \frac{0,80\ x\ 0,20}{30}\right)$

$$P(\hat{p}< 0,70)= P\left(Z < \frac{0,70-0,80}{\sqrt{\frac{0,80\ x\ 0,20}{30}}}\right) = P(Z < -1,37) = 0,0853.$$

Isso significa que a probabilidade de menos de 70% dos 30 indivíduos terem eficácia no tratamento de anemia ferropriva com suplementação de 50mg de ferro é de 8,53%.

Para resolução no software R, utilize o seguinte comando:

Onde além da função e argumentos já descritos anteriormente, também foi usado o argumento "sqrt" o qual calcula a raiz quadrada, neste caso do nosso desvio padrão colocado em parênteses.



16. Intervalo de Confiança

Alguns Conceitos:

Parâmetro: é um valor de resumo relacionado à variável de estudo considerando-se todos os elementos da população no cálculo, isto é, é uma medida de resumo populacional. Pode ser média, mediana, desvio padrão, variância ou proporção. Normalmente é inferido ou estimado por meio de uma amostra representativa da população.

Estimador: é uma função matemática que é calculada usando-se os valores da variável obtidos na amostra. Pode ser média amostral, variância amostral.

Estimativa: é o resultado obtido através da aplicação de um estimador.

Estimativa intervalar: É uma estimativa de um parâmetro a partir de um intervalo de valores, geralmente obtido com uma estimativa pontual e uma margem de erro subtraída e somada.

Intervalo de confiança: é a estimativa intervalar associada a um determinado nível de confiança geralmente expresso em porcentagem.

- ***Fórmula de intervalo de confiança para média*** μ ***(com variância populacional*** σ^2 ***desconhecida):***

$$\text{IC}(\mu;\ 100(1-\alpha)\%) = \left[\ \bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}};\ \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right],$$

em que $P(T_{n-1} > t_{\alpha/2}) = \alpha/2$ e T_{n-1} é uma variável aleatória com distribuição t-Student com $n - 1$ graus de liberdade.

- ***Fórmula de intervalo de confiança para média*** μ ***(com variância populacional*** σ^2 ***conhecida):***

$$\text{IC}(\mu;\ 100(1-\alpha)\%) = \left[\ \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}};\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

em que $P(Z > z_{\alpha/2}) = \alpha/2$ e Z é uma variável aleatória com distribuição normal padrão.

- ***Fórmula de intervalo de confiança para proporção p:***

$$\text{IC}(p;\ 100(1\text{-}\alpha)\%) = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\ };\ \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right],$$

em que $P(Z > z_{\alpha/2}) = \alpha/2$ e Z é uma variável aleatória com distribuição normal padrão.

- ***Fórmula de intervalo de confiança para variância σ^2:***

$$\text{IC}(\sigma^2;\ 100(1\text{-}\alpha)\%) = \left[\frac{(n-1)s^2}{Q_{\alpha/2}};\ \frac{(n-1)s^2}{Q_{1-\alpha/2}}\right],$$

em que $P(\chi^2_{n-1} > Q_{\alpha/2}) = \alpha/2$, $P(\chi^2_{n-1} > Q_{1-\alpha/2}) = 1 - \alpha/2$ e χ^2_{n-1} é uma variável aleatória com distribuição Qui-quadrado com $n - 1$ graus de liberdade.

O mesmo intervalo obtido para a variância σ^2 pode também ser utilizado para o desvio-padrão σ , bastando tirar as raízes dos extremos do intervalo,

$$\text{IC}(\sigma;\ 100(1\text{-}\alpha)\%) = \left[\sqrt{\frac{(n-1)s^2}{Q_{\alpha/2}}};\ \sqrt{\frac{(n-1)s^2}{Q_{1-\alpha/2}}}\right].$$

Exercícios:

16.1) Em uma pesquisa realizada com 25 praticantes de atividade física, a média de tempo semanal utilizado para realização de exercícios aeróbicos foi de 90 minutos, com desvio padrão de 20 minutos. Qual é a estimativa intervalar, a um nível de 90% de confiança, para o tempo médio de todos os praticantes de atividade física?

Observando o enunciado, podemos perceber que a variância populacional não nos é informada. Assim sendo, utilizamos a fórmula de intervalo de confiança para média populacional com variância desconhecida.

Como (1-α)100% = 90%, obtemos que α = 0,1 e $t_{0,05}$ = 1,7109 (obtido na tabela da distribuição t-Student ou no software R, tal que $P(T_{n-1} > t_{0,05}) = 0,05$). Desta forma:

Margem de erro: 1,7109 $\times \frac{20}{\sqrt{25}}$ = 6,84.

$IC(\mu;\ 90\%)$ = [90 ± 6,84] = [83,16 ; 96,84].

Ou seja, estima-se, com 90% de confiança, que a média de tempo de exercício aeróbico semanal dos praticantes de atividade física desta população esteja entre 83,16 e 96,84 minutos.

16.2) Em um experimento com uma amostra de n = 30 mulheres adultas, a média de cálcio consumido diariamente foi de 2000 mg com desvio padrão de 700 mg.

a) Qual o consumo médio de cálcio por essas mulheres com um intervalo de 95% de confiança?

Como (1-α)100% = 95%, obtemos que α = 0,05 e $t_{0,025}$ *= 2,0452 (obtido na tabela da distribuição t-Student ou no software R, tal que* $P(T_{n-1} > t_{0,025}) = 0,05$*. Desta forma:*

Margem de erro: 2,0452 $\times \frac{700}{\sqrt{30}}$ *= 261,38 mg.*

$IC(\mu;\ 95\%)$ *= [1738,62 ; 2261,38].*

Ou seja, estima-se que o consumo médio diário de cálcio pelas mulheres desta população, com 95% de confiança, esteja entre 1738,62 e 2261,38 mg.

b) Suponha que a amostra tenha sido maior, com n=46 mulheres. Neste caso, qual seria o valor da média amostral com esse mesmo intervalo de confiança? Suponha que o desvio padrão amostral tenha sido o mesmo.

Como (1-α)100% = 95%, obtemos que α = 0,05 e $t_{0,025}$ *= 2,0141 (obtido na tabela da distribuição t-Student ou no software R, tal que* $P(T_{n-1} > t_{0,025}) = 0,05$*. Desta forma:*

Margem de erro: 2,0141 $\times \frac{700}{\sqrt{46}}$ *= 207,87 mg.*

$IC(\mu;\ 95\%)$ *= [1792,12 ; 2207,87].*

Ou seja, estima-se que o consumo médio diário de cálcio pelas mulheres desta nova amostra, com 95% de confiança, esteja entre 1792,12 e 2207,87 mg.

16.3) Foi realizada uma pesquisa envolvendo uma amostra de 600 pacientes de um certo hospital. Cada um desses pacientes foi submetido a uma série de exames clínicos e, dentre outras coisas, mediu-se o Índice Cardíaco (em litros/min/m²) de todos eles. Os 600 pacientes foram então classificados, de forma aleatória, em 40 grupos de 15 pacientes cada. Para um desses grupos os valores medidos do Índice Cardíaco foram: 405, 348, 365, 291, 135, 260, 300, 155, 34, 294, 758, 472, 559, 143, 172.

a) Com base nos valores acima, construa um intervalo de confiança para o valor médio populacional do Índice Cardíaco ao nível de 95%.

Inicialmente, calcularemos a média amostral e o desvio padrão amostral da seguinte forma:

$\bar{x} = \frac{405+348+365+\ldots+143+172}{15} = 312,73$

$s = \sqrt{\frac{(405-312,73)^2+(348-312,73)^2+\ldots+(172-312,73)^2}{15-1}} = 185,80$

Como (1-α)100% = 95%, obtemos que α = 0,05 e $t_{0,025}$ = 2,1448 (obtido na tabela da distribuição t-Student ou no software R, tal que $P(T_{n-1} > t_{0,025}) = 0,05$). Desta forma:

Margem de erro: 2,1448 $\times \frac{185,80}{\sqrt{15}}$ = 102,89 L/min/m² ,

$IC(\mu;\ 95\%)$ = *[209,84 ; 415,62].*

Ou seja, estima-se que o índice cardíaco médio dos pacientes deste hospital, com 95% de confiança, esteja entre 209,84 e 415,62 L/min/m² .

b) Se para cada um desses 40 grupos de 15 pacientes fosse construído um intervalo de confiança para ao nível de 95%, quantos desses intervalos se espera que não conteriam a verdadeira média populacional no seu interior? Por que?

Como o valor de α adotado no caso foi 0,05 (1-95%), cerca de 5%, ou seja, 2 dos 40 (0,05 x 40 = 2) intervalos de confiança assim obtidos não conteriam em seu interior a verdadeira média populacional.

16.4) Deseja-se saber a eficácia de uma nova vacina contra determinada doença em grávidas, ou seja, deseja-se estimar proporção p de mulheres que não desenvolveriam a doença se previamente vacinadas. Uma amostra de 180 grávidas foi imunizada com a nova vacina e 110 delas não desenvolveram doença quando em contato com o vírus. Qual a proporção de imunização dessa vacina com um intervalo de confiança de 90%?

Inicialmente calcularemos a proporção amostral da seguinte forma:

Proporção = $\frac{110}{180}$ = *0,61.*

Como (1-α)100% = 90%, obtemos que α = 0,1 e $Z_{0,05}$ = 1,6449 (obtido na tabela da Normal ou no software R). Desta forma:

$IC(\mu;\ 90\%) = \left[0,61 - 1,6449\sqrt{\frac{0,61(0,39)}{180}};\ 0,61 + 1,6449\sqrt{\frac{0,61(0,39)}{180}}\right]$

= *[0,55 ; 0,67].*

Assim, com base nesta amostra, estimamos que a proporção de imunização de grávidas com a nova vacina está entre 55% e 67%, com 90% de confiança.

16.5) Um nutricionista, a fim de estimar o desvio padrão das medidas de um adipômetro (equipamento utilizado para medir a densidade corporal de um indivíduo) realizou um experimento. Um de seus pacientes, com densidade conhecida de 1,07, foi medido 10 vezes com esse adipômetro, obtendo-se um desvio padrão das medidas igual a 0,4. Supondo que as medidas do peso sigam uma distribuição normal, qual o intervalo de 90% de confiança para a variância das densidades?

Obtemos o intervalo para a variância populacional como:

$$IC(\sigma^2;90\%) = \left[\frac{(9)0{,}4^2}{Q[9;0{,}05]};\ \frac{(9)0{,}4^2}{Q[9;0{,}95]}\right] = \left[\frac{1{,}44}{16{,}919};\ \frac{1{,}44}{3{,}325}\right] = [0{,}085;\ 0{,}433].$$

Assim, o intervalo de 90% de confiança para desvio-padrão da densidade corporal é calculado como:

$$IC(\sigma;\ 90\%) = \left[\sqrt{0,085};\sqrt{0,433}\right] = [0{,}291;0{,}658].$$

Isso significa que, com base nesse experimento, estimamos que o desvio padrão das densidades corporais obtidas por esse adipômetro está entre 0,291 e 0,658, com 90% de confiança.

16.6) Em uma determinada região produtora de vinhos, foi criado um programa de regulamentação e fiscalização desses vinhos e de seus produtos derivados a fim de melhorar a qualidade dos produtos ofertados aos consumidores. Em uma dada amostra de 1000 produtos oriundos do vinho foram obtidos os dados de não conformidade expressos no quadro abaixo:

Número de não conformidade	0	1	2	3	4
Unidades	600	300	60	35	5

a) Qual a proporção de itens não conformes com um intervalo de confiança de 98%?

Inicialmente calcularemos a proporção amostral da seguinte forma:

$$Proporção = \frac{400}{1000} = 0{,}4.$$

Como (1-α)100% = 98%, obtemos que $\alpha = 0{,}02$ *e* $Z_{0{,}01} = 2{,}33$ *(obtido na tabela da Normal ou no software R). Desta forma:*

$$IC(\mu;\ 98\%) = \left[0,4 - 2,33\sqrt{\frac{0,4(0,6)}{1000}};\ 0,4 + 2,33\sqrt{\frac{0,4(0,6)}{1000}}\right]$$

$$= [0,36\ ;\ 0,44].$$

Assim, com base nesta amostra, estimamos que a proporção de produtos com inconformidades está entre 36% e 44%, com 98% de confiança.

16.7) O conteúdo de licopeno de tomates colhidos em uma determinada região, em miligramas, é uma variável aleatória que se supõe ter distribuição normal. Pretende-se obter a variabilidade do conteúdo de licopeno desses frutos colhidos ao longo da estação. Para isso, uma amostra de 11 tomates, de diferentes safras, foi avaliada e os teores de licopeno em miligramas, para cada 100g de tomate, obtidos foram os seguintes:
98 97 102 100 98 101 102 105 95 102 100

Construa um intervalo de confiança para a variância de unidades colhidas, com um grau de confiança igual a 95%.

Inicialmente, calcularemos a média de licopeno encontrada nos tomates dessa amostra, bem como o desvio padrão amostral

$$\overline{X} = \frac{98+97+102+100+98+101+102+105+95+102+100}{11} = 100$$

$$s^2 = \frac{(98-100)^2+(97-100)^2+\ldots+(100-100)^2}{11-1} = 8$$

Pela tabela da distribuição Qui-Quadrado com 10 graus de liberdade, temos que: $Q_{0,025} = 3,25$ e $Q_{0,975} = 20,48$.

$$IC(\sigma^2;95\%) = \left[\frac{(10)8}{Q[10;0,025]};\ \frac{(10)8}{Q[10;0,975]}\right] = \left[\frac{80}{3,25};\ \frac{80}{20,48}\right] = [3,90;\ 24,61].$$

Isso significa que, com base nesse experimento, estimamos que a variância do teor de licopeno presente nos tomates obtida neste estudo está entre 3,90 e 24,61 mg/100g, com 95% de confiança.

16.8) Uma pesquisa realizada com idosos de instituições de longa permanência para idosos (ILPI) na região centro sul de Belo Horizonte tinha como objetivo avaliar a presença de fragilidade nesse público. Para isso, 264 idosos foram avaliados com base na escala de fragilidade IVCF-20 e os resultados obtidos demonstraram de 58% desses idosos se enquadravam como “idosos frágeis”, enquanto 95 deles foram classificados como “idosos em risco de fragilização” e o restante como “idoso robusto”. Pede-se:

a) Qual a proporção de idosos frágeis com um intervalo de confiança de 95%?

Como o enunciado já revelou a proporção de idosos frágeis dessa amostra, p=0,58, podemos utilizar a seguinte fórmula para o cálculo:

Já que (1-α)100% = 95%, obtemos que α = 0,05 e $Z_{0,025}$ *= 1,96 (obtido na tabela da Normal ou no software R). Desta forma:*

$$IC(p;\ 95\%) = \left[0,58 - 1,96\sqrt{\frac{0,58(0,42)}{264}};\ 0,58 + 1,96\sqrt{\frac{0,58(0,42)}{264}}\right]$$

= [0,52 ; 0,64].

Assim, com base nesta amostra, estimamos que a proporção de idosos frágeis está entre 52% e 64%, com 95% de confiança.

b) Qual a proporção de idosos em risco de fragilização com um intervalo de confiança de 90%?

Inicialmente calcularemos a proporção amostral da seguinte forma:

Proporção = $\frac{95}{264}$ *= 0,36.*

Como (1-α)100% = 90%, obtemos que α = 0,1 e $Z_{0,05}$ *= 1,65 (obtido na tabela da Normal ou no software R). Desta forma:*

$$IC(p;\ 90\%) = \left[0,36 - 1,65\sqrt{\frac{0,36(0,64)}{264}};\ 0,36 + 1,65\sqrt{\frac{0,36(0,64)}{264}}\right]$$

= [0,31 ; 0,41].

Assim, com base nesta amostra, estimamos que a proporção de idosos em risco de fragilização está entre 31% e 41%, com 90% de confiança.

c) Qual a proporção de idosos robustos com um intervalo de confiança de 95%?

Inicialmente calcularemos a proporção amostral da seguinte forma:

Proporção de idosos robustos = 100% - (58% + 36%).

Proporção de idosos robustos = 6% = 0,06.

Já que (1-α)100% = 95%, obtemos que α = 0,05 e $Z_{0,025}$ = 1,96 (obtido na tabela da Normal ou no software R). Desta forma:

$$IC(p;\ 95\%) = \left[0,06 - 1,96\sqrt{\frac{0,06(0,94)}{264}};\ 0,06 + 1,96\sqrt{\frac{0,06(0,94)}{264}}\right]$$

$$= [0,03\ ;\ 0,09]\ .$$

Assim, com base nesta amostra, estimamos que a proporção de idosos robustos está entre 3% e 9%, com 95% de confiança.

16.9) Considere um profissional de educação física interessado em lançar um novo programa de musculação. Ele contrata uma empresa para realizar uma pesquisa de mercado com 500 pessoas. Entre as 500 pessoas, 157 manifestaram interesse em participar do programa. Faça um intervalo de confiança com 92% para a probabilidade de uma pessoa aderir ao referido programa.

O problema apresenta uma distribuição em que p é desconhecido. Assim, calcularemos inicialmente a proporção de pessoas interessadas em aderir ao programa:

$$Proporção = \frac{157}{500} = 0,314.$$

Como (1-α)100% = 92%, obtemos que α = 0,08 e $Z_{0,04}$ = 1,75 (obtido na tabela da Normal ou no software R). Desta forma:

$$IC(\mu;\ 92\%) = \left[0,314 - 1,75\sqrt{\frac{0,314(0,686)}{500}};\ 0,314 + 1,75\sqrt{\frac{0,314(0,686)}{500}}\right]$$

$$= [0,278\ ;\ 0,350]\ .$$

Assim, com base nesta amostra, estimamos que a probabilidade de uma pessoa aderir ao programa está entre 27,8% e 35%, com 92% de confiança.

16.10) Em um determinado hospital a hidratação de soro via venosa para pacientes com uma determinada condição de saúde pode ocorrer por dois métodos. São atendidos 50 pacientes pelo método A, que demora em média 10 minutos para o esvaziamento total do frasco de soro. Ao mesmo tempo, 70 pacientes são atendidos pelo método B, com esvaziamento médio de 12 minutos. Admitindo a variância do tempo igual a 100min² e distribuição normal dos tempos, calcule:

a) Um intervalo de confiança para o tempo médio do esvaziamento do soro feito pelo método A, considerando nível de confiança de 92%.

Organizando os dados temos que μ_A *= 10 min e* n_A *= 50.*

Como α *= (1- 0,92) = 0,08,* $Z_{0,04}$ *= 1,75 (obtido na tabela da Normal ou no software R). Desta forma:*

$$IC(\mu;\ 92\%) = \left[10 - 1,75\sqrt{\frac{100}{50}};\ 10 + 1,75\sqrt{\frac{100}{50}}\right] = [7,5\ ;\ 12,5].$$

Ou seja, o tempo médio de esvaziamento do soro pelo método A é de 7,5 a 12,5 minutos com um nível de confiança de 92%.

b) Faça o mesmo para o método B.

Organizando os dados temos que μ_B *= 12 min e* η_B *= 70.*

Como α*=0,08,* $Z_{0,04}$*= 1,75 (obtido na tabela da Normal ou no software R). Desta forma:*

$$IC(\mu;\ 92\%) = \left[12 - 1,75\sqrt{\frac{100}{70}};\ 10 + 1,75\sqrt{\frac{100}{70}}\right] = [9,9\ ;\ 14,1].$$

Ou seja, o tempo médio de esvaziamento do soro pelo método B é de 9,9 a 14,1 minutos com um nível de confiança de 92%.

c) Calcular o I.C. para a diferença entre os tempos médios de esvaziamento entre os métodos A e B

Organizando os dados temos que μ_A *= 10 min e* n_A *= 50 e* μ_B*= 12 min e* n_B *= 70*

$$IC(\mu;\ 92\%) = \left[(10 - 12) - 1,75\sqrt{100(\frac{1}{70} + \frac{1}{50})};\ 10 + 1,75\sqrt{100(\frac{1}{70} + \frac{1}{50})}\right]$$

$$= [-5,24\ ;\ 1,24].$$

Isso significa, que a diferença entre os tempos médios de esvaziamento do método A para o método B estão entre -5,24 minutos a 1,24 minutos, com um intervalo de confiança de 92% para ambos. Ou seja, o método A demora de 5,24 minutos a menos que até 1,24 minutos a mais que o método B.

17. Teste de Hipóteses para Médias e Proporções

Alguns Conceitos:

Hipótese: é uma afirmação sobre parâmetros de uma ou mais populações.

Teste de hipóteses: é o processo de decisão entre duas hipóteses sobre um ou mais parâmetros de uma ou mais populações.

Hipótese nula: Uma hipótese que no procedimento de teste é assumida verdadeira e a partir da amostra observada decidimos se esta hipótese deve ser rejeitada ou não. Sempre é expressa contendo uma igualdade.

Hipótese alternativa: é a afirmação escolhida quando a hipótese nula é considerada falsa.

Exercícios:

17.1) Identifique a hipótese nula e alternativa para cada situação abaixo:

a) Um biólogo estudando uma espécie de pássaros de sua região observou que no verão a média diária de vôo é de 3km. Porém, ele suspeita que no inverno essa média seja menor, para isso, deseja realizar um estudo com esses pássaros durante essa estação.

Seja μ representando a média diária de vôo, em quilômetros, então as hipóteses de interesse são:

$$H_o: \mu \geq 3km \text{ versus } H_1: \mu < 3km \text{ ou } H_o: \mu = 3km \text{ versus } H_1: \mu < 3km.$$

b) Em um hospital, a ocorrência de desnutrição em pacientes idosos internados é de 70% para os que possuem algum tipo de doença crônica. Num grupo de 200 idosos desse hospital, 130 foram diagnosticados com desnutrição. Devido aos novos métodos adotados no hospital buscando a menor prevalência de desnutrição em idosos, suspeita-se que essa proporção tenha diminuído.

Seja p representando a prevalência de desnutrição em idosos, então as hipóteses de interesse são:

$$H_o: p \geq 0,7 \text{ versus } H_1: p < 0,7 \text{ ou } H_o: p = 0,7 \text{ versus } H_1: p < 0,7.$$

c) Um instituto de produção de vacinas afirma que em todas as ampolas devem ter 7ml da vacina. Porém, suspeita-se que houve problemas na produção de um determinado lote e que, neste lote, a quantidade média nas ampolas está diferente da afirmada. Para verificar

essa suspeita, o instituto selecionou uma amostra aleatória de 70 ampolas desse lote, observando uma quantidade média de vacina igual a 6.4 mg e um desvio padrão de 0.6 mg.

Seja μ representando a quantidade média de vacina na ampola, em ml, então as hipóteses de interesse são:

$$H_o: \mu = 7ml \text{ versus } H_1: \mu \neq 7ml.$$

d) A associação médica do estado de Minas Gerais está muito preocupada com o caso de suicídios entre a classe, cuja média, nos últimos tempos, tem sido de 26 suicídios por ano com desvio padrão de 2 casos. Tentou-se um programa de prevenção ao suicídio, buscando sua redução, após o qual foi coletado dados anuais de suicídio em 15 clínicas médicas distribuídas no estado e observou-se o número de 3 suicídios.

Seja μ representando a média de suicídios entre a classe médica, então as hipóteses de interesse são:

$$H_o: \mu \geq 26 \text{ versus } H_1: \mu < 26 \text{ ou } H_o: \mu = 26 \text{ versus } H_1: \mu < 26.$$

e) O consumidor de um certo produto acusou o fabricante, dizendo que mais de 20% das unidades fabricadas apresentam defeito. Para confirmar sua acusação, ele usou uma amostra de tamanho 50, onde 27% dos produtos eram defeituosos.

Seja p representando a prevalência de defeitos dos produtos deste fabricante, então as hipóteses de interesse são:

$$H_o: p \leq 0{,}2 \text{ versus } H_1: p > 0{,}2 \text{ ou } H_o: p = 0{,}2 \text{ versus } H_1: p > 0{,}2.$$

Teste de hipótese para a média populacional com variância desconhecida

17.2.1) Um instituto de produção de vacinas afirma que em todas as ampolas devem ter 7ml da vacina. Porém, suspeita-se que houve problemas na produção de um determinado lote e que, nesse lote, a quantidade média nas ampolas está diferente da afirmada. Para verificar essa suspeita, o instituto selecionou uma amostra aleatória de 61 ampolas desse lote, observando uma quantidade média de vacina igual a 6,4 ml e um desvio padrão de 0,6 ml. Assuma um nível de significância de 0,01.

Neste caso, as hipóteses são:

$$H_o: \mu = 7ml \text{ versus } H_1: \mu \neq 7ml,$$

em que μ representa a média das vacinas em cada ampola, em ml.

Como se trata de um teste para uma média populacional em que a variância populacional σ^2 é desconhecida, o valor observado para a estatística de teste é dado por:

$t_0 = \frac{6,4-7}{0,6/\sqrt{61}} = -7,81.$

Para tomar uma decisão, podemos calcular o p-valor através de:

valor-p = $P(T_{60} < -7,81) = P(T_{60} > 7,81) < 0,005 < \alpha = 0,01$.

Portanto, visto que o valor-p é menor do que o nível de significância, rejeitamos a hipótese nula.

Rejeitamos H_o ao nível de 1% de significância, ou seja, há evidências amostrais de que o nível médio de vacina nas ampolas daquele lote não é 7ml.

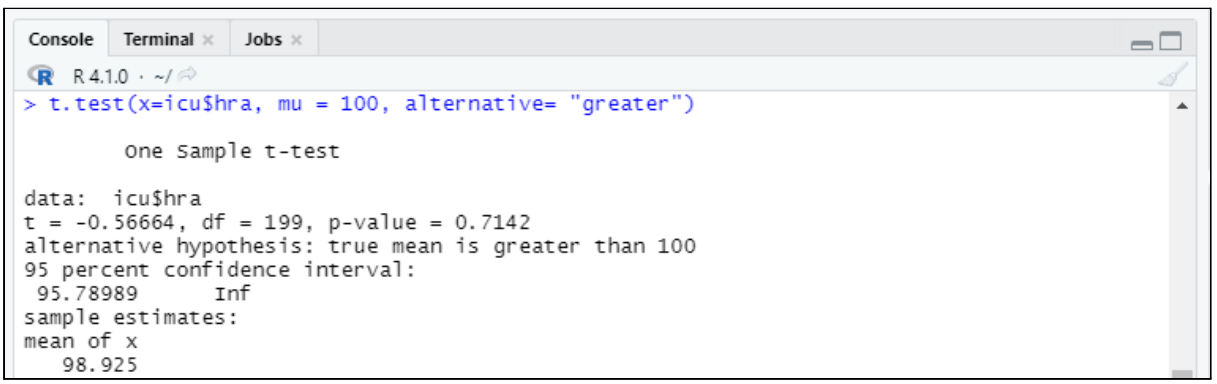
Teste de hipótese para a média populacional com variância desconhecida - Software R

17.2.2) **(Banco de dados icu - Pacote Aplore 3)** Um dos funcionários da equipe de UTI acredita que a média da frequência cardíaca dos indivíduos na admissão está bastante elevada, acreditando que a é média maior do que 100 batimentos por minuto (bpm) (variável "hra"). Suponha que se esse funcionário estiver correto a equipe deverá ter uma atenção especial a esse parâmetro. A um nível de 5% de significância, a equipe deve aceitar a hipótese desse funcionário?

Seja μ representando a média da frequência cardíaca dos indivíduos. Neste caso, as hipóteses são:

H_o: μ = 100 bpm versus H_1: $\mu > 100$ bpm ou H_o: $\mu \leq 100$ bpm versus H_1: $\mu > 100$ bpm

Para resolução no software utilize a função "t.test", que é usada para realizar os testes de hipótese que utilizam da tabela t-Student. O argumento "mu" se refere ao valor testado da média e o argumento "alternative" deve ser utilizado para indicar se a hipótese alternativa é maior ("greater") ou menor ("less") que a hipótese nula.



Sendo assim, visto que o valor-p é maior do que o nível de significância, não rejeitamos a hipótese nula. A um nível de 5% de significância, não rejeitamos H_o, logo não há evidências amostrais de que o número médio de bpm dos pacientes na admissão na UTI seja maior que 100 bpm.

Teste de hipótese para a média populacional com variância conhecida

17.3.1) Numa determinada clínica de fisioterapia, avalia-se sistematicamente a força manual dos clientes por meio de um dinamômetro. Dessas avaliações, sabe-se que a média de força do braço direito de homens adultos dessa clínica é de 53 FPM e a variância 16. Após a realização de um novo programa de forças com esses pacientes, deseja-se verificar se houve alguma melhora. Uma amostra de 15 adultos obteve média igual a 50 FMP. Qual é a conclusão ao nível de significância de 5 %?

Neste caso, as hipóteses são:

$$H_o: \mu = 53 \text{ FPM versus } H_1: \mu \neq 53 \text{ FPM},$$

em que μ representa a média de força no braço direito de homens adultos.

O valor observado da estatística de teste é $Z_{obs} = \frac{(x-\mu)\sqrt{n}}{\sigma} = \frac{(50-53)\sqrt{15}}{\sqrt{16}} = -2{,}90.$

Solução pela abordagem do p-valor: Usando a tabela normal padrão, encontramos área na cauda superior após 2,90 igual a 0,0019. Como o teste é bilateral, o resultado é o dobro deste valor, ou seja (p = 2 x 0,0019 = 0,0038) que é menor que $\alpha = 0,05$. Portanto, rejeitamos a hipótese nula.

Ao nível de 5% de significância, há evidências de que o novo programa de força tenha gerado alguma alteração nos resultados obtidos por esses homens adultos.

Teste de hipótese para a média populacional com variância conhecida - Software R

17.3.2) **(Banco de dados icu - Pacote Aplore 3)** Suponha que uma diretriz de saúde recomenda que hospitais que possuem pacientes na UTI (Unidade de Tratamento Intensivo), com pressão sistólica média acima de 130 mmHg, priorizem a compra de um medicamento X que em estudos se mostrou mais seguro para esses pacientes, apesar de apresentar um valor mais elevado. A um nível de 5% de significância, a média da pressão sistólica dessa amostra, indica a compra desse medicamento X? Considere que essa amostra possua variância igual a 18.

Neste caso, as hipóteses são:

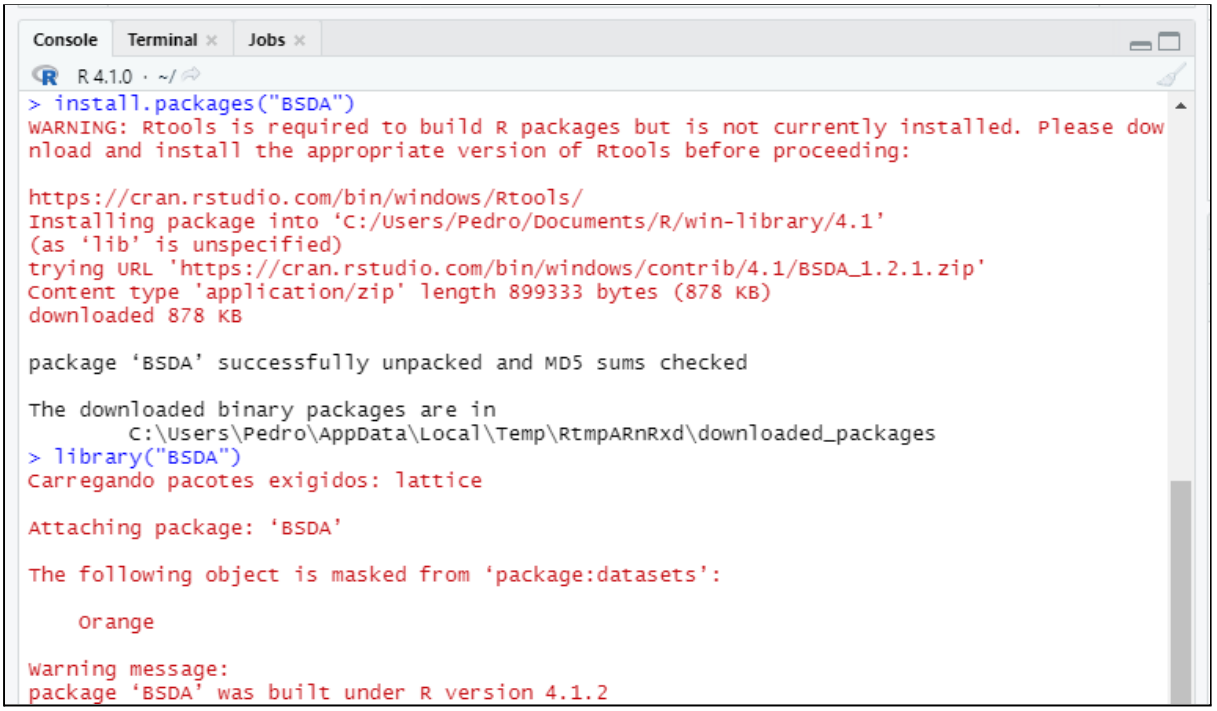
H_o: $\mu = 130$ mmHg versus H_1: $\mu > 130$ mmHg

ou

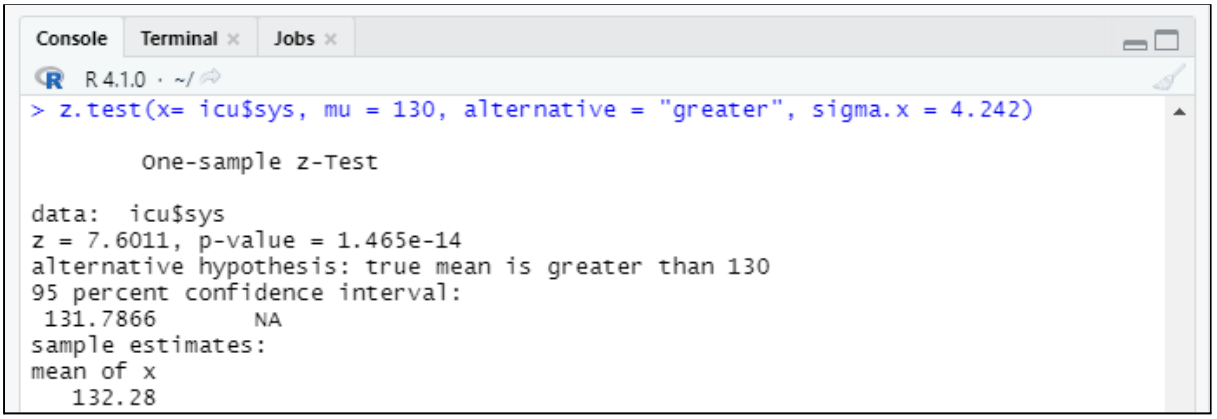
H_o: $\mu \leq 130$ mmHg versus H_1: $\mu > 130$ mmHg,

em que μ representa a pressão sistólica média populacional.

Primeiramente, para os testes com variância conhecida, que utilizam os dados de uma distribuição normal padrão, será necessário instalar um pacote adicional ao R (veja figura abaixo), que possui a função "z.test" destinada a esse tipo de teste de hipótese.



Feito isso, utilizamos o código abaixo para solucionar a questão. Para essa função, existe um argumento adicional, "sigma.x" onde deve-se indicar o desvio padrão da população de onde foi retirada a amostra.



Como o p- valor é menor do que 0,05, rejeitamos H_o ao nível de 5% de significância, ou seja, há evidências para recomendar que esse hospital compre o medicamento X para os pacientes de sua UTI.

Teste de hipótese para proporção com 1 população

17.4) Um obstetra afirma que menos de 10% dos recém nascidos na cidade em que ele trabalha, possuem baixo peso. Em um hospital dessa mesma cidade, em uma amostra de 60 recém nascidos, 3 deles possuíam baixo peso. Considerando um nível de significância de 5%, existe evidência de que a afirmação do obstetra está correta?

Neste caso as hipóteses são:

$$H_o: p = 0{,}10 \text{ versus } H_1: p < 0{,}10 \text{ ou } H_o: p \geq 0{,}10 \text{ versus } H_1: p < 0{,}10,$$

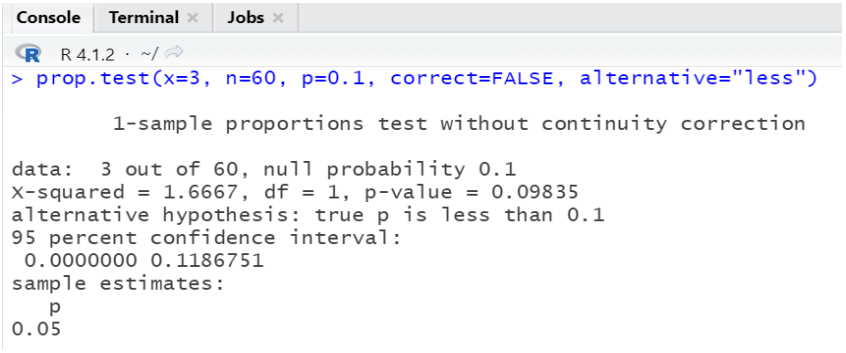
com p representando a proporção populacional de recém nascidos com baixo peso. O valor observado da estatística de teste é

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\ (1-p_0)}{n}}} = \frac{0{,}05 - 0{,}10}{\sqrt{\frac{0{,}10\ (1-0{,}10)}{60}}} = -1.2909.$$

O valor crítico a um nível de 5% de significância é de -1,645, separando os valores abaixo como região crítica/região de rejeição,assim sendo, como o valor do Z_{obs} é -1,29, ele não está na região crítica, não rejeitamos a hipótese nula. Assim sendo, não há evidências de que a afirmação do obstetra esteja correta com relação à proporção de recém nascidos com baixo peso.

A função "prop.test" presente no R é capaz de solucionar esta questão. Utilize o argumento "x" para descrever o número de eventos, "n" para amostra total, "p" para a proporção da hipótese, e o argumento "correct= FALSE" para indicar que não deve ser feita nenhuma correção de continuidade (o que é feito quando aproxima-se uma variável aleatória discreta por uma variável aleatória contínua).

Observação: o motivo da estatística de teste no R ser maior é devido ao fato de que o software utiliza a distribuição Qui-quadrado ao invés da distribuição normal.



Teste de hipótese para amostras dependentes teste T pareado

17.5) Um grupo de pesquisa tem a hipótese de que a adição de 10g de fibras solúveis em um determinada refeição conseguiria diminuir a glicemia pós-prandial (estado alimentado). Foi realizado um experimento com 10 indivíduos, os quais em um dia mediram sua glicemia após essa refeição, e em outro dia mediram a glicemia consumindo a mesma refeição mais a ingestão de 10g de fibras solúveis. Com um nível de significância de 10%, existem evidências para sustentar a hipótese da pesquisa?

Glicemia pós prandial sem fibra	Glicemia pós prandial com fibras	d	d^2
140	130	10	100
132	126	6	36
135	135	0	0
120	103	17	289
110	107	3	9
105	100	5	25
127	112	15	225
121	118	3	9
110	109	1	1
123	115	8	64
		Σ =68	Σ =758

Nessa situação trabalhamos com a média populacional (μ_d) das diferenças entre o antes e depois, e portanto, as hipóteses são:

$$H_0: \mu_d \leq 0 \quad versus \quad H_1: \mu_d > 0.$$

Deve-se calcular a média das diferenças e o desvio padrão das diferenças:

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{68}{10} = 6{,}8 \text{ e } S_d = \sqrt{\frac{\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}}{n-1}} = \sqrt{\frac{758 - \frac{(68)^2}{10}}{10-1}} =$$

5,73.

Com isto, calculamos o valor observado da estatística de teste: $t_{obs} = \frac{\bar{d} - \mu_d}{\frac{S_d}{\sqrt{n}}} =$

$$\frac{6{,}8 - 0}{\frac{5{,}73}{\sqrt{10}}} = 3{,}757.$$

O valor crítico a um nível de significância de 10% na tabela t-Student é de 1,383, delimitando os valores acima como região crítica/região de rejeição. Já que t_{obs} *é igual a 3,757 que é maior que 1,383, ele está dentro da região crítica, assim rejeitamos* H_0 *. Assim, podemos concluir que, com 10% de significância, há evidências a favor de que 10g de fibras solúveis podem reduzir os níveis de glicemia pós prandial.*

A resolução dessa questão utilizando o software R seria da seguinte maneira: Primeiramente, crie variáveis com os valores de glicemia dados na questão. As amostras devem ser colocadas em ordem de acontecimento. O argumento "pared = TRUE" indica que as amostras são pareadas.

```
Console   Terminal   Jobs
R 4.1.2 · ~/
> glicemiasemfibra = c(140, 132, 135, 120, 110, 105, 127, 121, 110, 123)
> glicemiacomfibra = c(130, 126, 135, 103, 107, 100, 112, 118, 109, 115)
> t.test(glicemiasemfibra, glicemiacomfibra, paired = TRUE, alternative = "greater")

        Paired t-test

data:  glicemiasemfibra and glicemiacomfibra
t = 3.7521, df = 9, p-value = 0.00227
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 3.477843      Inf
sample estimates:
mean of the differences
                    6.8
```

Teste de hipótese para amostras independentes e variâncias conhecidas

17.6.1) Uma pesquisadora em saúde pública afirma que o volume de um determinado tumor, quando detectado, em média, é maior nos Estados Unidos do que no Brasil. A justificativa é pelo diagnóstico do sistema de saúde pública. Em pesquisas anteriores encontrou-se que para essas populações o desvio padrão do volume desses tumores é de 36 ml para o Estados Unidos e de 32 ml para o Brasil. Uma pesquisa foi realizada estudando o tamanho médio desse tumor, nos Estados Unidos uma amostra de 50 tumores resultou em um volume médio de 300 ml e no Brasil uma amostra de 50 pacientes obteve um volume médio de 280 ml . A partir dessas informações, em um nível de significância de 1% existem evidências indicando que a afirmação da pesquisadora está correta?

Considere μ_1 o volume médio do tumor nos Estados Unidos e μ_2 o volume médio no Brasil. Nesse caso as hipóteses são:

$$H_0: \mu_1 = \mu_2 \text{ versus } H_1: \mu_1 > \mu_2 \text{ ou } H_0: \mu_1 \leq \mu_2 \text{ versus } H_1: \mu_1 > \mu_2.$$

Calculando a estatística de teste, temos:

$$z_{obs} = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(300-280) - 0}{\sqrt{\frac{36^2}{50} + \frac{32^2}{50}}} = 2{,}94.$$

O valor crítico para 1% de significância é de 2,33, delimitando os valores acima como região crítica/região de rejeição, sendo o valor de z_{obs} 2,94, maior do que 2,33, isso significa que ele se encontra na região crítica, portanto não rejeitamos a hipótese nula.

A um nível de 1% de significância, existem evidências para suportar a afirmação da pesquisadora de que o volume de tumores do determinado tumor, quando detectado, é maior nos Estados Unidos do que no Brasil.

Teste de hipótese para amostras independentes e variâncias conhecidas - Software R

17.6.2) Um grupo de biólogos estudam tartarugas em extinção, e um dos biólogos acredita que a média de ovos postos pela espécie B é maior do que a média de ovos postos pela espécie A. Sabe-se por estudos anteriores que a variância de ovos postos pelas espécies é de 10 na espécie A e 12 na B. Sendo assim, a um nível de 5% de significância, podemos aceitar a hipótese do biólogo? Considere que as amostras abaixo para o teste são suficientemente grande.

espécie_A = c(77, 80, 64, 67, 78, 68, 63, 73, 73, 85, 82, 61, 73, 71, 76, 82, 65, 71, 80, 76, 65, 69, 64, 69, 73, 72)

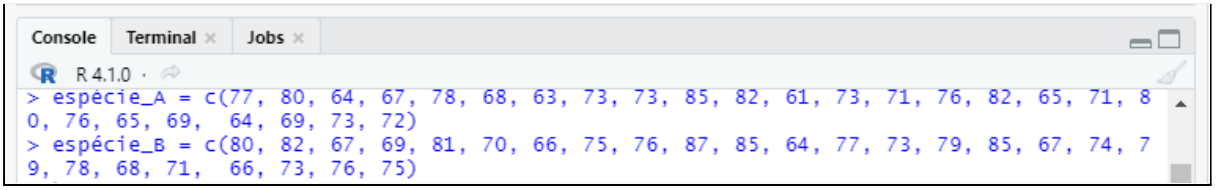
espécie_B = c(80, 82, 67, 69, 81, 70, 66, 75, 76, 87, 85, 64, 77, 73, 79, 85, 67, 74, 79, 78, 68, 71, 66, 73, 76, 75)

Para este caso as hipóteses são:

$$H_0: \mu_A = \mu_B \text{ versus } H_1: \mu_A < \mu_B \text{ ou } H_0: \mu_A \geq \mu_B \text{ versus } H_1: \mu_A < \mu_B,$$

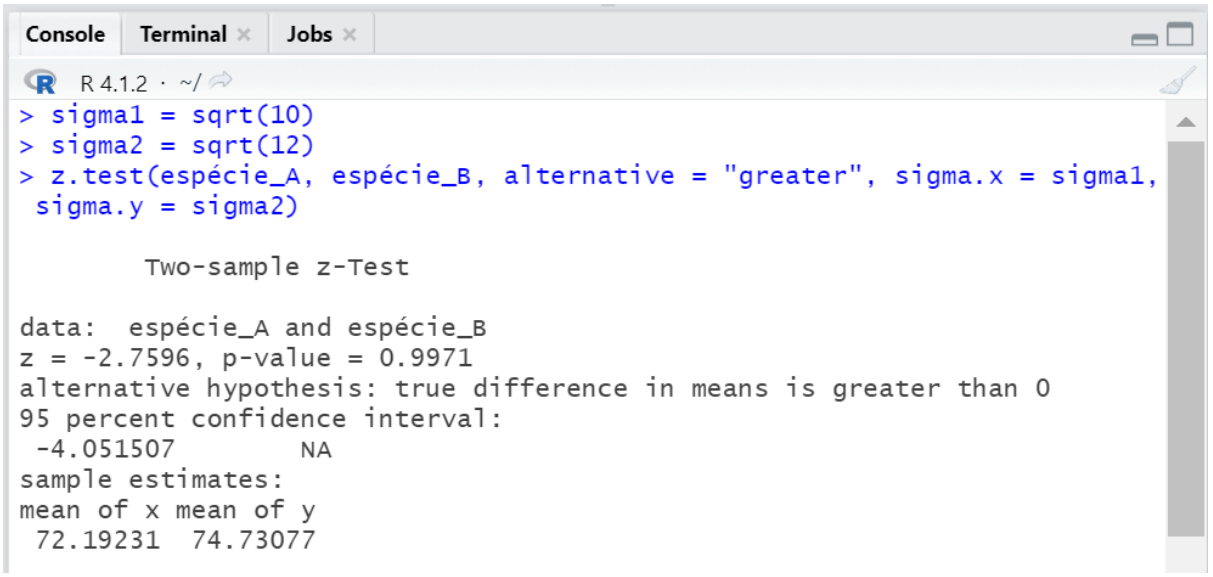
em que μ_A e μ_B são a média de ovos postos pela espécie A e espécie B, respectivamente.

Inicialmente insira as amostras em objetos do R:



Utilize a função a seguir:

Perceba que as amostras são escritas em ordem, bem como a variância de cada amostra, portanto, o "sigma.x" é referente ao desvio padrão da amostra "espécie_A" e o "sigma.y" é referente ao desvio padrão da amostra "espécie_B" .



Como o p-valor é maior do que 0,05, a um nível de 5% de significância, não existem evidências para apoiar a afirmação do biólogo de que em média as tartarugas da espécie B põem mais ovos do que as tartarugas da espécie A.

Teste de hipótese para amostras independentes e variâncias desconhecidas e iguais

17.7.1) Pesquisadores realizaram uma investigação para avaliar os efeitos da utilização de ventosaterapia na diminuição da dor muscular em atletas. Dois grupos de atletas foram questionados sobre o nível de dor, em uma escala de 0 a 10, sendo que o grupo 1 foi submetido à terapia com ventosas e o grupo 2 não recebeu nenhuma intervenção.

Escores de dor:

Grupo 1	$n_1 = 45$	$\overline{X}_1 = 6,75$	s1 = 1,59
Grupo 2	$n_2 = 46$	$\overline{X}_2 = 6,98$	s2 = 1, 74

Considerando um nível de significância de 1%, teste a hipótese dos pesquisadores de que a utilização de ventosaterapia contribui para reduzir a dor muscular em atletas.

As hipóteses são:

H_0: $\mu_1 = \mu_2$ (A ventosaterapia, em média, não contribui para reduzir a dor).

versus

H_1: $\mu_1 < \mu_2$ (A ventosaterapia, em média, contribui para reduzir a dor),

em que μ_1 e μ_2 representam os níveis médios de dor com e sem ventosaterapia. Devemos calcular a estatística de teste:

$$t_{obs} = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(6,75 - 6,98) - 0}{\sqrt{\frac{1,59^2}{45} + \frac{1,74^2}{46}}} = \frac{(-0,23)}{\sqrt{0,056 + 0,065}}$$

$$= -0,663.$$

Definição do grau de liberdade: menor valor entre $n_1 - 1$ e $n_2 - 1$, logo, temos que o grau de liberdade será 44.

Definição do valor crítico na tabela t student: como em H_1: $\mu_1 < \mu_2$, o valor crítico será o simétrico do valor obtido na tabela t a um nível de significância 0,01 e um grau de liberdade igual a 44. Logo, esse valor será - 2,414, que delimita uma região de rejeição abaixo dele. Como a estatística de teste - 0,663 não está na zona de rejeição de H_o, devemos aceitar a hipótese nula (H_o).

Teste de hipótese para amostras independentes e variâncias desconhecidas e iguais - Software R

17.7.2) Suponha que um nutricionista queira testar, a um nível de 5% de significância, a eficácia de um novo suplemento no mercado que diz aumentar o desempenho em exercícios de curta duração e alta intensidade, como por exemplo a corrida de 100 metros, portanto, os atletas realizariam esse percurso em um tempo reduzido. Para isso, são selecionados dois grupos de corredores de clubes diferentes, de mesma variabilidade, com 30 corredores cada, considerando que as variâncias desconhecidas são iguais. O grupo A não utilizou esse suplemento, e o grupo B utilizou conforme orientação do fabricante. Os tempos em segundos dos grupos foram os seguintes:

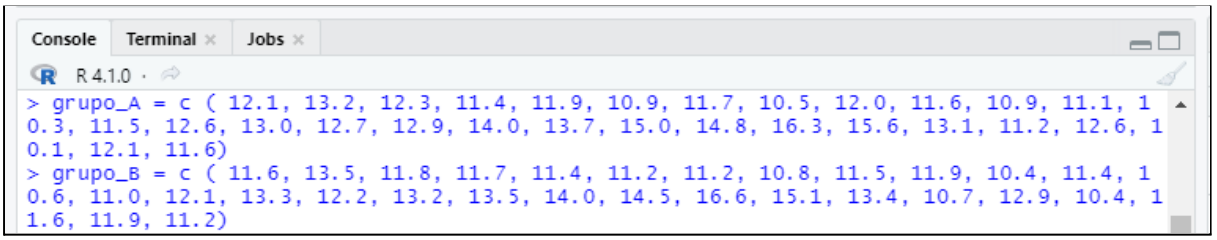
grupo_A = c (12.1, 13.2, 12.3, 11.4, 11.9, 10.9, 11.7, 10.5, 12.0, 11.6, 10.9, 11.1, 10.3, 11.5, 12.6, 13.0, 12.7, 12.9, 14.0, 13.7, 15.0, 14.8, 16.3, 15.6, 13.1, 11.2, 12.6, 10.1, 12.1, 11.6)

grupo_B = c (11.6, 13.5, 11.8, 11.7, 11.4, 11.2, 11.2, 10.8, 11.5, 11.9, 10.4, 11.4, 10.6, 11.0, 12.1, 13.3, 12.2, 13.2, 13.5, 14.0, 14.5, 16.6, 15.1, 13.4, 10.7, 12.9, 10.4, 11.6, 11.9, 11.2)

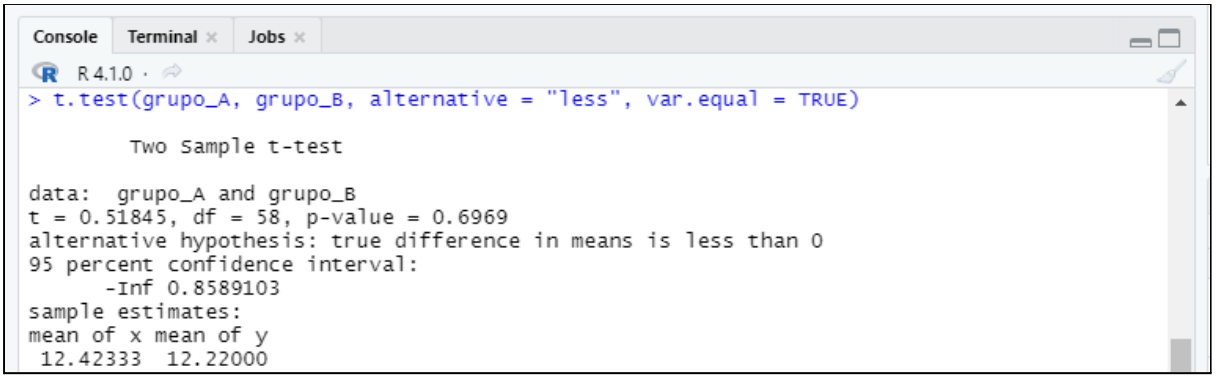
Neste caso as hipóteses são:

$$H_0: \mu_1 = \mu_2 \text{ versus } H_1: \mu_1 > \mu_2 \text{ ou } H_0: \mu_1 \geq \mu_2 \text{ versus } H_1: \mu_1 > \mu_2,$$

em que μ_1 e μ_2 representam os tempos médios entre os atletas que não utilizam e utilizam o suplemento, respectivamente. Para resolução no programa, inicialmente inserimos os valores das amostras no software.



Em seguida digite o código abaixo, em que o argumento "var.equal = TRUE" definirá que as variâncias de ambas as amostras são iguais.



Sendo assim, visto que o valor-p é maior do que o nível de significância, não rejeitamos a hipótese nula.

A um nível de 5% de significância, não rejeitamos H_0, pois não há evidências amostrais de que o tempo médio gasto para correr 100 metros pelos percorridos pelos atletas do grupo B, seja menor do que o tempo gasto pelos atletas do grupo A.

Teste de hipótese para amostras independentes e variâncias desconhecidas e diferentes

17.8.1) Sabe-se que a estimulação elétrica funcional (FES) pode ser usada na terapia de fortalecimento muscular em indivíduos com alguma sequela de AVC. Para se avaliar os efeitos dessa terapia, pesquisadores analisaram o aumento médio da força muscular em dois grupos de indivíduos com sequelas de AVC, sendo que o grupo 1 foi submetido ao tratamento e o grupo 2 não recebeu nenhuma intervenção.

Grupo 1	$n_1 = 15$	$\bar{x}_1 = 37,2\ kgf$	$s_1 = 2,27$
Grupo 2	$n_2 = 16$	$\bar{x}_2 = 35,8\ kgf$	$s_2 = 2,18$

Considere as variâncias populacionais são iguais ($\sigma_1 = \sigma_2$) e um nível de significância de 0,05. Teste a hipótese de que o grupo submetido à estimulação elétrica funcional (FES) teve um aumento médio de força muscular maior que o grupo não submetido à intervenção.

Temos as seguintes hipóteses:

H_0: $\mu_1 = \mu_2$ (A FES não contribui para aumentar a força muscular).

versus

H_1: $\mu_1 > \mu_2$ (A FES contribui para aumentar a força muscular),

em que μ_1 e μ_2 representam o aumento médio de forma muscular com e sem estimulação elétrica funcional, respectivamente.

Cálculo da variância amostral combinada (Sp^2) :

$$Sp^2 = \frac{(n_1 - 1)\times s_1^{\,2} + (n_2 - 1)\times s_2^{\,2}}{n_1 + n_2 - 2} = \frac{(15-1)\times 2{,}27^2 + (16-1)\times 2{,}18^2}{15+16-2}$$

$$= \frac{14\times 5{,}15 + 15\times 4{,}75}{29} = 4{,}94.$$

Cálculo da estatística de teste:

$$t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_{p^2}}{n_1} + \frac{s_{p^2}}{n_2}}} = \frac{37{,}2 - 35{,}8 - 0}{\sqrt{\frac{4{,}94}{15} + \frac{4{,}94}{16}}} = \frac{1{,}4}{\sqrt{0{,}33+0{,}30}} = 1{,}77.$$

Determinação do grau de liberdade(G.L.):

G.L = $n_1 + n_2 - 2$ = 31-2 = 29.

Determinação do valor crítico na tabela t-Student, como em H_1: $\mu_1 > \mu_2$, o valor crítico será obtido na tabela t a um nível de significância 0,05 com grau de liberdade igual a 29. Logo, o valor crítico será 1,699, que delimita uma região acima dele onde será rejeitada a hipótese nula (H_0.).

Como a estatística de teste 1,77, conclui-se que a um nível de 5% de significância, a FES contribui para aumentar a força muscular.

Teste de hipótese para amostras independentes e variâncias desconhecidas e diferentes - Software R

17.8.2) Um farmacêutico estuda um antitérmico (medicamento para febre), e quer saber se a um nível de significância de 5%, o grupo intervenção, tratado com antitérmico (grupo_B), terá temperatura média menor do que o grupo controle, tratado com placebo, (grupo_A). Os grupos haviam 35 participantes cada, possuem variabilidade diferente, os valores de febre medidos foram os seguintes:

grupo_A = c (40.5, 39.3, 39.9, 40.1, 39.8, 41.1, 40.3, 40.2, 41.1, 39.9, 40.5, 39.7, 40.2, 40.1, 40.2, 40.7, 39.9, 39.9, 40.5, 41.2, 39.8, 41.1, 39.9, 40.3, 40.8, 40.6, 40.9, 39.5, 39.6, 40.9, 39.3, 39.6, 40.2, 40.4, 39.6)

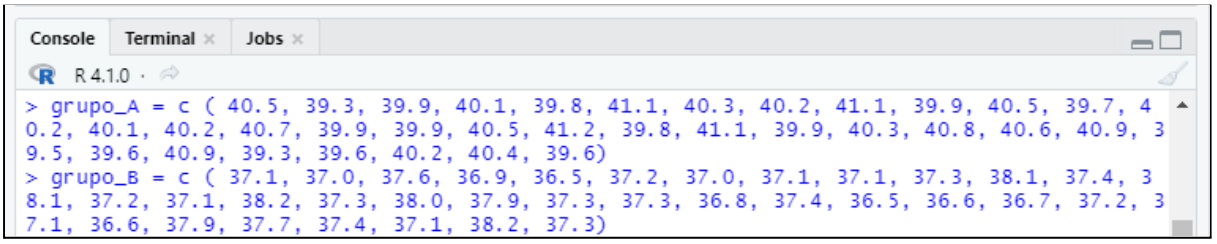
grupo_B = c (37.1, 37.0, 37.6, 36.9, 36.5, 37.2, 37.0, 37.1, 37.1, 37.3, 38.1, 37.4, 38.1, 37.2, 37.1, 38.2, 37.3, 38.0, 37.9, 37.3, 37.3, 36.8, 37.4, 36.5, 36.6, 36.7, 37.2, 37.1, 36.6, 37.9, 37.7, 37.4, 37.1, 38.2, 37.3)

As hipóteses nessa situação são:

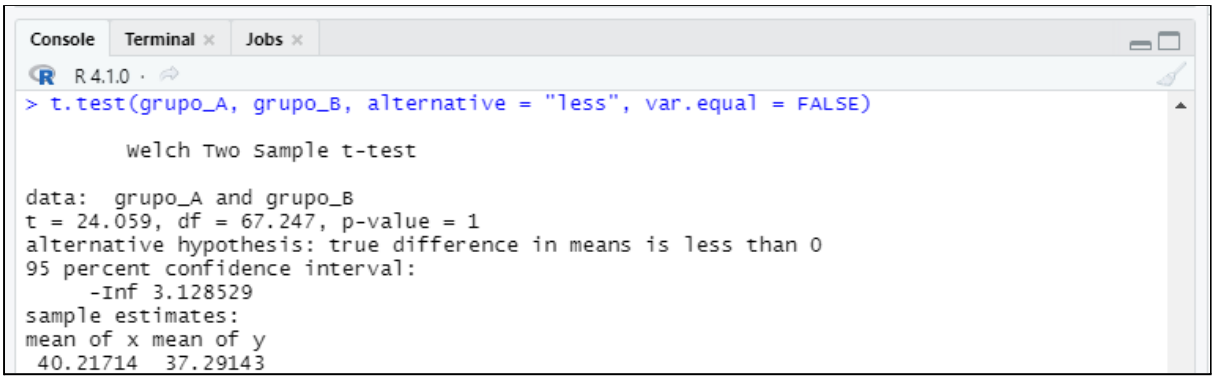
$$H_0: \mu_1 = \mu_2 \text{ versus } H_1: \mu_1 > \mu_2 \text{ ou } H_0: \mu_1 \leq \mu_2 \text{ versus } H_1: \mu_1 > \mu_2,$$

em que μ_1 e μ_2 representam as temperaturas médias das pessoas que tomam placebo e antitérmico, respectivamente.

As amostras devem ser computadas pelo software:



Com as amostras salvas, utilize o código abaixo, em que o argumento "var.equal = FALSE" irá identificar que as variâncias das amostras são diferentes.



Sendo assim, visto que o valor-p é maior do que o nível de significância, não rejeitamos a hipótese nula.

A um nível de 5% de significância, não há evidências amostrais de que o medicamento reduziu a febre dos pacientes do grupo_B em comparação aos do grupo_A.

Teste de hipótese para igualdade de proporções de 2 populações

17.9) Um fisioterapeuta tem o hábito de incluir alongamentos antes de iniciar uma sequência de exercícios em seus pacientes. Então, resolveu investigar a influência desses alongamentos na amplitude de movimento (ADM) realizada durante a execução da tarefa. Foram analisados dois grupos, um que não fez alongamentos (grupo 1) e outro que fez alongamentos antes das atividades (grupo 2). O grupo 1 apresentava 13 indivíduos, dos quais 6 apresentaram ganho de amplitude, enquanto que o grupo 2 com 12 indivíduos, dos quais 7 apresentaram ganho nesse desfecho. Para realizar essa investigação, o profissional buscou comparar as taxas de ganho de ADM entre os dois grupos de pacientes, para verificar se havia diferença significativa entre as taxas de ganho de amplitude, adotando um nível de significância de 0,05.

Temos as seguintes hipóteses:

H_0: $p_1 = p_2$ *(Não há diferença entre as taxas de ganho de ADM).*

versus

H_1: $p_1 \neq p_2$ *(Há diferença entre as taxas de ganho de ADM),*

em que p_1 *e* p_2 *representam proporção de pessoas com ganho de amplitude sem e com alongamentos, respectivamente.*

Para calcular a estatística de teste calculamos a proporção amostral ponderada, que representa a estimativa supondo que as duas populações têm a mesma proporção:

$\bar{p} = \frac{y_1 + y_2}{n_1 + n_2}$ *, sendo* y_1 *e* y_2 *o número de sucessos observados (também obtível por* $n_i \times p_i$*).*

Considerando que $p_1 = \frac{6}{13} = 0,46$ *;* $p_2 = \frac{7}{12} = 0,58$ *e* $\bar{p} = \frac{6+7}{13+12} = \frac{13}{25} = 0,52$.

O valor observado da estatística de teste é dado por:

$$Z_{obs} = \frac{p_1 - p_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{0,46 - 0,58}{\sqrt{\frac{0,52 \times 0,48}{13} + \frac{0,52 \times 0,48}{12}}} = \frac{-0,12}{\sqrt{1,92 + 2,08}} = -0,06.$$

Como se trata de um teste bilateral, temos que:

p-valor = 2P ($Z \geq |Z_{0bs}|$) = 2×0,4761 = 0,952 (nível de significância), não rejeitamos H_0.

Como o p-valor > α (nível de significância), não rejeitamos H_0.

Teste de hipótese para igualdade de proporções de 2 populações - Software R

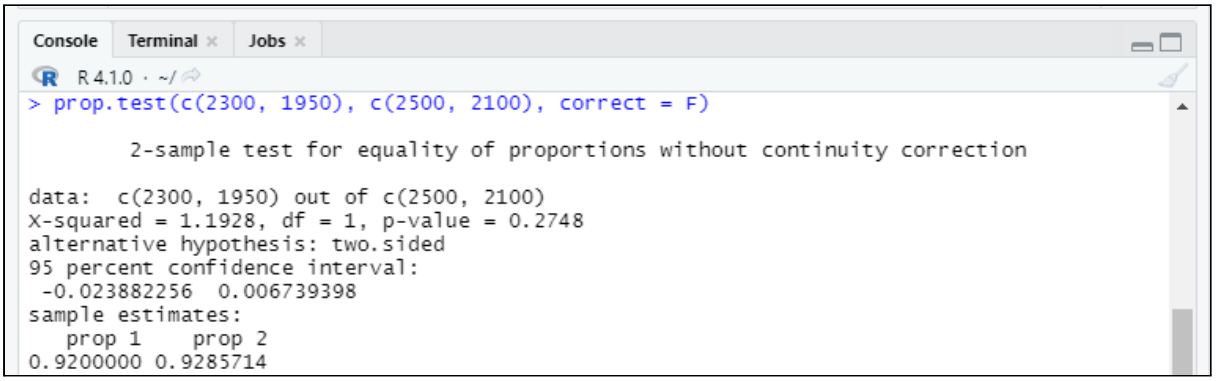
17.9) O serviço de saúde pública de uma determinada cidade quer saber saber se existe diferença na adesão a vacinação contra covid-19, dependendo da vacina ofertada. Para isso são comparados dois bairros que receberam tipos de vacinas diferentes, o bairro A recebeu a vacina A, sendo que dos 2500 cidadãos, 2300 se vacinaram, já o bairro B recebeu a vacina B, e dos 2100 cidadãos, 1950 se vacinaram. A um nível de significância de 5%, houve diferença entre a proporção de vacinados entre os bairros dessa cidade?

As hipóteses a serem testadas são:

$$H_0: p_1 = p_2 \text{ versus } H_1: p_1 \neq p_2 ,$$

em que p_1 e p_2 representam as proporções de adesão a vacinação com as vacinas A e B, respectivamente.

No software a função necessária para resolução dessa questão é a seguinte:



Pelo fato de que o valor-p é maior do que o nível de significância, não rejeitamos a hipótese nula.

A um nível de 5% de significância, não rejeitamos H_0, pois não há evidências amostrais de que a proporção da adesão a vacinação seja entre a vacina A e a vacina B nos bairros avaliados.

18. Teste Qui-Quadrado

Alguns Conceitos:

Definição: O teste Qui-Quadrado de Pearson é utilizado para avaliar três tipos de objetivos: bondade do ajuste (também chamado teste de aderência), homogeneidade e independência. O teste para bondade do ajuste foge do escopo deste material, e no tocante aos outros dois testes, temos:

- **Teste de Homogeneidade:** Verifica se as frequências de indivíduos em categorias distintas de uma variável seguem o mesmo padrão em diferentes subgrupos definidos segundo outra variável.
    - Ex: Escolha de atividades - faculdades, viagens, empregos - de alunos formados do ensino médio depois de um ano de formados, para ver se o número de alunos escolhendo uma certa atividade tem mudado de classe para classe, ou de década para década.
- **Teste de Independência:** Avalia se as respostas para uma variável categórica são independentes das respostas de outra variável categórica avaliadas na mesma unidade amostral.
    - Ex: Avaliar se o voto das pessoas em uma eleição depende da nacionalidade de cada pessoa.

Suposições:

- As observações são independentes;
- Os itens de cada grupo são selecionados aleatoriamente;
- As variáveis são categóricas e os dados observados são organizados em uma tabela com frequências;
- Cada observação pertence a uma e somente uma categoria;
- A frequência esperada em cada casela da tabela deve ser maior ou igual a 5.

Aplicações:

1. Avaliar se há uma relação de dependência entre duas variáveis categóricas;
2. Avaliar se há homogeneidade de proporções das respostas de uma variável categórica para em diferentes subgrupos definidos por uma segunda variável categórica.

No caso 1, as hipóteses são:

H_0: As variável X_1 e X_2 são independentes;
H_1: As variável X_1 e X_2 são dependentes.

Já no caso 2, as hipóteses são:

H_0: As proporções das respostas são homogêneas para todas as subpopulações;
H_1: Ao menos uma subpopulação tem distribuição de proporções diferentes das demais.

Estatística de Teste: A estatística de teste é dada por:

$$\chi^2_{obs} = \sum_{i=1}^{N_c} \frac{(O_i - E_i)^2}{E_i},$$

em que:

- N_c: Número de caselas na tabela (Número de Linhas × Número de Colunas);
- $i = 1, 2, 3, \cdots, N_c$, representando cada casela da tabela;
- O_i: Valor observado na casela "i";
- E_i: Valor esperado na casela "i", dado por $E_i = \frac{(total\ da\ linha) \times (total\ da\ coluna)}{(total\ da\ tabela)}$.

O grau de liberdade $g.l$, necessário para a observação do valor crítico na tabela, é dado por

$$g.l = (número\ de\ linhas - 1) \times (número\ de\ colunas - 1).$$

Obs.: A fórmula da estatística de teste também pode ser reescrita com dois somatórios para considerar linhas e colunas.

Exercícios:

18.1) Um pesquisador deseja saber se três espécies de pinguins são homogêneas quanto à proporção de indivíduos de cada sexo. Para obter a resposta, foram colhidas aleatoriamente amostras de cada espécie, de tamanhos variados.

As amostras continham 200 pinguins-imperadores, 180 pinguins-rei e 240 pinguins-macaroni. Observou-se que 125 dos pinguins-imperadores, 100 dos pinguins-reis e 146 dos pinguins-macaroni eram machos.

Os dados do experimento apresentam evidências estatísticas suficientes contra a hipótese de que as três espécies possuem a mesma proporção entre os sexos?

Observamos que temos três amostras de diferentes tamanhos, colhidas aleatoriamente, e desejamos realizar um teste de hipóteses para igualdade de proporções. Portanto, o teste utilizado será o Teste Qui-Quadrado.

Chamando as espécies de I (pinguins-imperadores), R (pinguins-rei) e M (pinguins-macaroni), temos as seguintes hipóteses:

H_0: *Todas as espécies têm proporções de machos e fêmeas iguais.*

versus

H_1: *Pelo menos uma das espécies tem proporção diferente das demais.*

Organizando os dados, podemos construir a seguinte tabela:

Tabela da Frequência de Pinguins em cada Espécie por Sexo			
Espécie	**Macho**	**Fêmea**	**Total Linha**
Imperador	125	75	200
Rei	100	80	180
Macaroni	146	94	240
Total Coluna	371	249	620

Sabendo que os valores esperados são obtidos através dos totais das linhas multiplicados pelos totais da coluna e divididos pelo total da tabela, podemos incrementar essa tabela, adicionando os valores esperados entre parênteses:

Tabela da Frequência de Pinguins em cada Espécie por Sexo com Frequências Esperadas			
Espécie	**Macho**	**Fêmea**	**Total Linha**
Imperador	125 (119,6774)	75 (80,3226)	200
Rei	100 (107,7097)	80 (72,2903)	180
Macaroni	146 (143,6129)	94 (96,3871)	240
Total Coluna	371	249	620

Para encontramos a estatística de teste χ^2_{obs}*, realizamos a seguinte operação:*

$$\chi^2_{obs} = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(125 - 119{,}6774)^2}{119{,}6774} + \frac{(100 - 107{,}7097)^2}{107{,}7097} + \ldots + \frac{(94 - 96{,}3871)^2}{96{,}3871} \approx 2{,}0623\,.$$

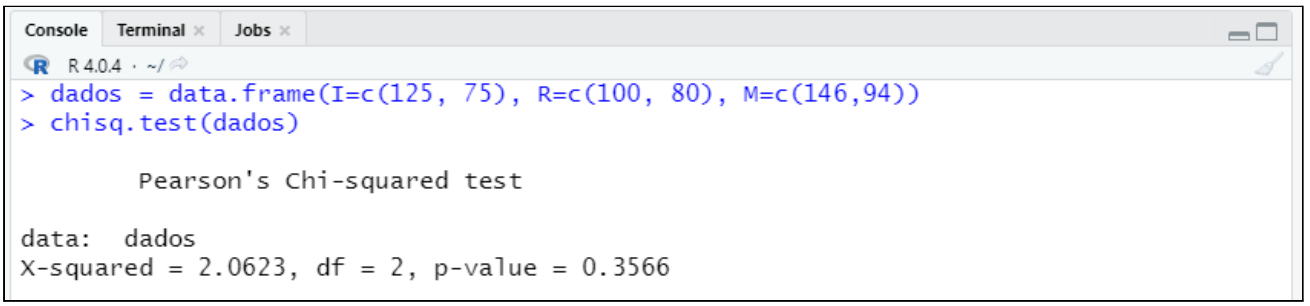
Os graus de liberdade são dados por:

$g.l$ = *(número de linhas - 1)* $\times$ *(número de colunas -1)* $= (3-1) \times (2-1) = 2.$

Portanto, devemos observar na tabela Qui-Quadrado, com dois graus de liberdade, a probabilidade acumulada à direita do valor observado, $P[\chi^2_2 > 2.0623]$*, que é maior que*

0.25, portanto maior que 0.05. Assim sendo, com $\alpha = 0.05$, não encontramos evidências para rejeitarmos H_0.

Para se solucionar esse problema através do R, basta entrar com os dados em formato "data.frame" e utilizar a função "chisq.test()", como mostrado a seguir:



18.2) Uma clínica de oftalmologia acredita que existem mais mulheres com miopia do que homens, sugerindo que o sexo biológico influencia na presença da doença. Para testar essa hipótese foi montado uma tabela com os dados dos pacientes atendidos por essa clínica no último ano.

Tabela de Frequências Observadas de Miopía por Sexo

Sexo	Miopia Presente	Miopia Ausente	Total Linha
Masculino	35	281	316
Feminino	46	256	302
Total Coluna	81	537	618

A um nível de significância de 10%, existe evidência para apoiar a hipótese dessa clínica?

Para este problema, as hipóteses são:

H_0: A presença de miopia independe do sexo biológico do indivíduo

versus

H_1: A presença de miopia depende do sexo biológico do indivíduo.

Observando a tabela com os dados coletados, nota-se que 81 pessoas com miopia representam 13,12% da amostra total e 537 pessoas sem miopia representam 86,89% da amostra total.

Sob hipótese de independência, esse mesmo percentual deveria ser encontrado para ambos os sexos, o que poderia ser ilustrado com uma tabela de valores esperados, veja:

Tabela esperada:

Tabela de Frequências Esperadas de Miopía por Sexo			
Sexo	Miopia Presente	Miopia Ausente	Total Linha
Masculino	41	275	316
Feminino	40	262	302
Total Coluna	81	537	618

Assim, temos que o valor observado da estatística de teste é:

$$\chi^2_{obs} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{(o_{11}-e_{11})^2}{e_{11}} + \frac{(o_{12}-e_{12})^2}{e_{12}} + \frac{(o_{21}-e_{21})^2}{e_{21}} + \frac{(o_{22}-e_{22})^2}{e_{22}}$$

$$= \frac{(35-41)^2}{41} + \frac{(281-275)^2}{275} + \frac{(46-40)^2}{40} + \frac{(256-262)^2}{262} = 0{,}88 + 0{,}13 + 0{,}9 + 0{,}14 = 2{,}05.$$

O valor crítico para 10% de significância com 1 grau de liberdade é de 2,706, delimitando os valores acima como região crítica/região de rejeição .

Pode-se concluir que, a um nível de significância de 10%, o valor observado da estatística de teste não se encontra na região de rejeição. Sendo assim, não há evidência para rejeitar a hipótese nula, ou seja, a presença de miopia independe do sexo do indivíduo.

18.3) **(Banco de dados lowbwt - Pacote Aplore3)** Um pediatra buscando fortalecer sua orientação às suas pacientes, deseja saber se o baixo peso de nascimento de um bebê (variável “low”) depende do hábito de fumar durante a gravidez (variável “smoke”). Considere um nível de significância de 5%.

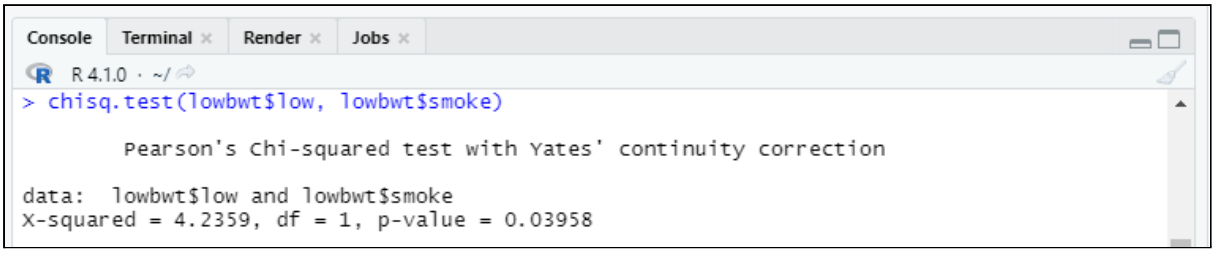
Para este problema, as hipóteses de interesse são:

H_0: O baixo peso de nascimento de um bebê independe do hábito de fumar durante a gravidez

versus

H_1: O baixo peso de nascimento de um bebê depende do hábito de fumar durante a gravidez.

Para solucionar essa questão basta utilizar a função chisq.test, do software R, utilizada para realização de testes Qui-Quadrado:



Sendo assim, visto o p-valor e considerando um nível de significância de 5%, rejeitamos a hipótese nula, ou seja, o baixo peso de nascimento de um bebê depende do hábito de fumar durante a gravidez.

18.4) **(Banco de dados icu - Pacote Aplore3)** Suponha que você faz parte da equipe de profissionais onde foi coletada a amostra, e precisa saber quais são as variáveis dependentes ou independentes em relação ao desfecho do paciente ao receber alta (se sobreviveu ou não). Realize testes de independência com as seguintes variáveis e descubra qual é a relação ao desfecho do paciente (variável "sta").

a) Apresentar provável infecção na admissão na UTI (variável "inf") possui dependência com o desfecho de alta (vida/morte) do paciente a um nível de 1% de significância?

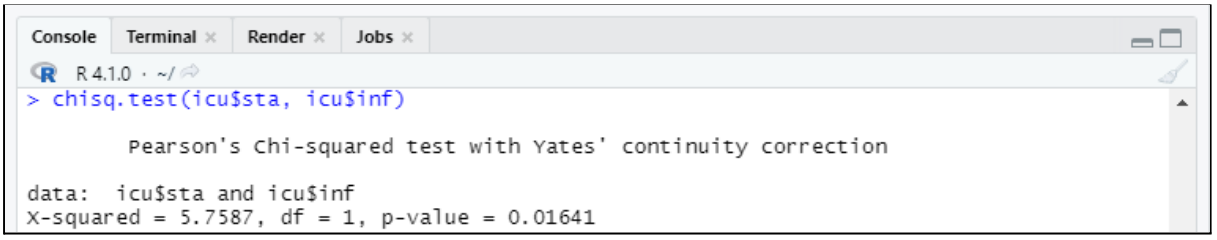
Para este problema, as hipóteses são:

H_0: O desfecho do paciente é independente de que exista uma provável infecção na admissão

versus

H_1: O desfecho do paciente é dependente de que exista uma provável infecção na admissão.

Para solucionar, basta usar o código abaixo:



Concluindo, observando o p-valor 0.016 > 1%, não rejeitamos a hipótese nula a um nível de 1% de significância, ou seja, o desfecho do paciente é independente de que exista uma provável infecção na admissão.

b) O tipo de admissão, eletiva ou emergencial (variável "type") é dependente do desfecho de alta (vida/morte) do paciente a um nível de 1% de significância?

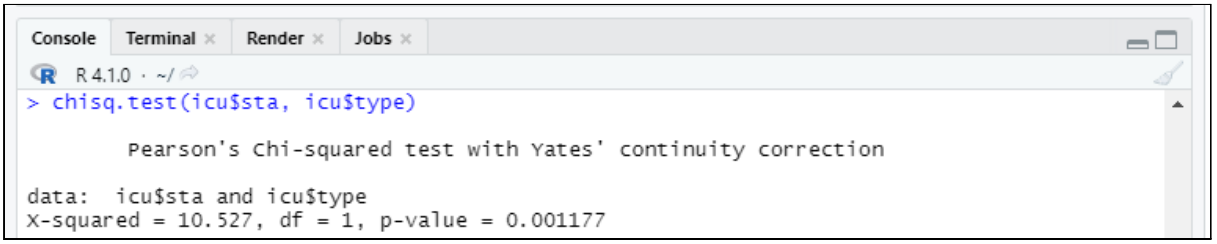
Neste caso, as hipóteses de interesse são:

H_0: *O desfecho do paciente é independente do tipo de admissão que o paciente teve*

versus

H_1: *O desfecho do paciente é dependente do tipo de admissão que o paciente teve.*

Os seguintes argumentos devem ser utilizados para a função "chisq.test":



A partir do resultado do p-valor, rejeitamos a hipótese nula a um nível de 1% de significância, sendo assim, o desfecho do paciente é dependente do tipo de admissão, se esta foi eletiva ou emergencial.

c) Pacientes que passaram por reanimação cardiorrespiratória (variável "cpr") possui maior dependência com o desfecho de alta (vida/morte) do paciente a um nível de 1% de significância?

As hipóteses são:

H_0: *O desfecho do paciente é independente de que o paciente tenha passado por reanimação cardiorrespiratória*

versus

H_1: *O desfecho do paciente é dependente de que o paciente tenha passado por reanimação cardiorrespiratória.*

Utilize o seguinte código:

```
> chisq.test(icu$sta, icu$cpr)

        Pearson's Chi-squared test with Yates' continuity correction

data:  icu$sta and icu$cpr
X-squared = 7.8209, df = 1, p-value = 0.005165

Warning message:
In chisq.test(icu$sta, icu$cpr) :
  Chi-squared approximation may be incorrect
```

Nesse caso não é possível tirar uma conclusão do teste, uma vez que o aviso presente no console do R Studio se refere a uma aproximação que é realizada automaticamente quando os dados para formulação da tabela não atende uma suposição necessária do teste Qui-Quadrado, de que todos os valores das caselas da tabela esperada, sejam maiores que 5. Sendo assim ,o resultado encontrado não é confiável.

Para confirmar solicite as tabelas utilizadas no teste Qui-Quadrado através dos seguintes códigos:

Tabela observada:

```
> chisq.test(icu$sta, icu$crn)$observed
        icu$crn
icu$sta  No Yes
  Lived 149  11
  Died   32   8
Warning message:
In chisq.test(icu$sta, icu$crn) :
  Chi-squared approximation may be incorrect
```

Tabela observada: Tabela esperada:

```
> chisq.test(icu$sta, icu$crn)$expected
        icu$crn
icu$sta    No  Yes
  Lived 144.8 15.2
  Died   36.2  3.8
Warning message:
In chisq.test(icu$sta, icu$crn) :
  Chi-squared approximation may be incorrect
```