ESCOLA SUPERIOR DE TECNOLOGIA DA SAÚDE DE LISBOA
INSTITUTO POLITÉCNICO DE LISBOA

Health Data Science

Pós-Graduação Health Data Science
Ano Letivo 2023/2024

Relatório – Análise Base de Dados

Adesão à terapia de hipertensão arterial em pacientes hipertensos recém tratados em unidades de cuidados de saúde primários

---

Discentes:
Inês Silva, nº 2023133
Maria Raquel Quintão, nº 2023134

Docente: Professora Carina Silva

Índice

Enquadramento 2

Enunciado 2

Seleção da amostra 3

Questões 5

1. Classifique as variáveis da BD (tipo e escala de medição) 5

2. Caracterize as variáveis Idade, Sexo, grupo_terap, nr_receita recorrendo a gráficos e/ou estatísticas de acordo com as características das variáveis e analise os resultados. 6

Idade 6

Sexo 8

Grupo_terap 9

Nr_receita 10

3. Recorrendo a gráficos, compare o tempo até à primeira aquisição (tempo_inicio) por grupo terapêutico (grupo_terap) 12

4. Obtenha uma representação gráfica que permita comparar a compra de medicamentos genéricos ou marca (marca_generico) por sexo e interprete-o. 15

5. Construa um diagrama de dispersão entre o poder de compra (Poder-compra) e o custo inicial e conclua. Complemente com uma estatística que considere adequada 17

6. Verifique se existe associação entre a variável "inicio" e o código ICPC. 20

7. Analise as variáveis descontinua, persist_6m, persist_24m, descontinua2, recorrendo a estatísticas e/ou gráficos e conclua sobre a adesão à terapêutica dos pacientes. 21

8. Obtenha uma estimativa pontual para o valor médio do poder de compra (Poder_compra) e uma estimativa intervalar com uma confiança de 99% apenas para os indivíduos que aviam dois ou mais medicamentos. 23

9. Compare a estimativa por intervalo de confiança a 95% para o mesmo parâmetro da alínea anterior e conclua. 26

10. Considerando apenas os indivíduos com idade inferior ou igual a 45 anos, pode afirmar-se que o valor médio dos valores obtidos para o custo inicial (Custo_inicial) é significativamente superior a 3? 27

11. Verifique se existem diferenças significativas entre o valor médio do poder de compra (Poder_compra) entre homens e mulheres. 30

12. Utilizando a base de dados de que dispõe proceda a uma análise estatística suplementar elaborando 3 questões e tire conclusões. 31

12.1. Compare o custo médio da primeira receita entre tipos de hipertensão e represente-o graficamente. 31

12.2 Verifique se existem diferenças significativas entre o valor médio do custo da primeira receita e o custo da receita passada 24 meses depois 32

12.3 Verifique se existe associação entre o sexo e o início do tratamento e tire conclusões. 34

Referências Bibliográficas 36

Enquadramento

No contexto da disciplina de Desenho de Investigação em Organizações de Saúde, inserida no programa de estudos da Pós-Graduação em Health Data Science da Escola Superior de Tecnologia da Saúde de Lisboa, foi atribuída a tarefa de realizar um trabalho em grupo que permitisse a aplicação dos conhecimentos estatísticos adquiridos, utilizando recursos informáticos, nomeadamente o R Studio.

Para a condução da análise estatística, foi fornecida uma base de dados no formato CSV, contendo 509 registos, os quais abrangem informações pertinentes à adesão à terapia de hipertensão arterial (AHT) por parte de pacientes recentemente diagnosticados com hipertensão, atendidos em unidades de cuidados de saúde primários.

Enunciado

A base de dados disponibilizada apresenta informação relativa à adesão à terapia de hipertensão arterial (AHT) em pacientes hipertensos recém-tratados em unidades de cuidados de saúde primários. Foram identificados todos os pacientes que foram diagnosticados com hipertensão e receberam uma primeira prescrição (prescrição de índice) para pelo menos um medicamento AHT entre março e abril de 2019.

Variáveis da base de dados:

1. Idade
2. Sexo: masculino / feminino
3. inicio: Doente inicia tratamento (sim ou não)
4. ICPC: código da hipertensão: K86 – hipertensão sem complicações / K87 – hipertensão com complicações
5. tempo_inicio: Tempo até à primeira aquisição do medicamento
6. grupo_terap: Grupo terapêutico. As classes de drogas AHT avaliadas, com os códigos do sistema de classificação química terapêutica anatómica (QTA) correspondentes foram:
    - 1 = C02: Anti-hipertensivos
    - 2 = C03: Diuréticos
    - 3 = C07: agentes bloqueadores beta
    - 4 = C08: bloqueadores do canal de cálcio
    - 5 = C09: Agentes que atuam no sistema renina-angiotensina

- 6 = 2 ou mais

7. nr_receita: número de medicamentos na 1ª receita (1-um; 2-dois ou mais)
8. marca_generico: comprou a marca ou o genérico (0-genérico; 1-marca)
9. Poder_compra: poder de compra (dados segundo a PORDATA)
10. Custo_inicial: custo da 1ª receita
11. Descontinua: O doente descontinua prematuramente: 0=não, 1=sim
12. Persist_6m: persistência ao fim de 6 meses: 0=não, 1=sim
13. Persits_24m: persistência ao fim de 24 meses: 0=não, 1=sim
14. Descontinua2: doente descontinua ao fim de 2 anos: 0=não, 1=sim
15. Custo_24: custo da receita passada 24 meses depois

Observação: o valor 999 na BD corresponde a um valor omisso. A análise deverá ter em consideração a remoção destes valores, pois caso contrário irão ser considerados nos diferentes cálculos. No R funções como na.rm = TRUE, na.omit(), servem para lidar com os valores omissos, porém no R os valores omissos são identificáveis através de NA.

Seleção da amostra

*Selecione aleatoriamente 20% dos indivíduos da BD BD_Epoca_Recurso_DIOS_2024.xls e guarde a nova DB num ficheiro *.xls. Esta será a nova BD que irão usar (…).*

- **População alvo:** todos os pacientes que foram diagnosticados com hipertensão e receberam uma primeira prescrição (prescrição de índice) para pelo menos um medicamento AHT entre março e abril de 2019, tratados em unidades de cuidados de saúde primários.
- **Amostra:** subconjunto representativo da população alvo, aleatório
- **Dimensão da amostra** (n): 101 (20% da população alvo)
- **Unidade estatística:** paciente/doente

Nas indicações práticas prestadas, foi-nos pedido que selecionássemos aleatoriamente 20% dos indivíduos da BD fornecida. Para tal, carregámos o ficheiro original disponibilizado em .xls no R – Import Dataset > From Excel. Na importação, atribuímos a NA o valor “999”. Calculámos o valor correspondente a 20% da população e criámos o dataset “amostra” através do uso da função “sample()”, criando uma amostra aleatória do vetor “BD_original”.

Posteriormente, após a instalação do pacote “openxlsx”, guardámos o ficheiro “amostra_bd.xlsx”.

Para sabermos em que variáveis existem valores omissos, para posteriormente podermos ter em conta a existência dos mesmos na análise e resolução dos exercícios, utilizámos a função *colSums(is.na(amostra))* que retorna o total de valores NA por coluna/variável da amostra.

```
      idade        sexo      inicio        ICPC  time_inicio  grupo_terap
          0           0           0          24           12           0
   nr_medic marca_generico poder_compra custo_inicial  descontinua  persist_6m
          0          18           0           0           12          12
persist_24m  descontinua2     custo_24
         12          12           0
```

Assim, verificamos que as variáveis ICPC, time_inicio, marca_generico, descontinua, persist_6m, persist_24m e descontinua2 apresentam valores omissos.

Questões

1. Classifique as variáveis da BD (tipo e escala de medição).

Tabela 1 *Classificação das variáveis da BD*

Variáveis	Tipo	Escala de medição
Idade	Quantitativa discreta	Razão
Sexo	Qualitativa	Nominal
Inicio	Qualitativa	Nominal
ICPC	Qualitativa	Nominal
tempo_inicio	Quantitativa contínua	Razão
grupo_terap	Qualitativa	Nominal
nr_receita	Qualitativa	Ordinal
marca_generico	Qualitativa	Nominal
Poder_compra	Quantitativa contínua	Razão
Custo_inicial	Quantitativa contínua	Razão
Descontinua	Qualitativa	Nominal
Persist_6m	Qualitativa	Nominal
Persits_24m	Qualitativa	Nominal
Descontinua2	Qualitativa	Nominal
Custo_24	Quantitativa contínua	Razão

2. Caracterize as variáveis Idade, Sexo, grupo_terap, nr_receita recorrendo a gráficos e/ou estatísticas de acordo com as características das variáveis e analise os resultados.

Idade

A variável **idade** é quantitativa discreta, pelo que é importante analisar a sua distribuição para melhor a caracterizar. Nesse sentido, calculámos a média, mediana, moda, mínimo, máximo, bem como o coeficiente de assimetria (skewness) e curtose (kurtosis), com recurso ao software R.

Tabela 2 *Estatísticas da variável "Idade"*

Min	1º Q	Mediana	Média	3º Q	AIQ	Max	Skew	Kurt	Desvio Padrão	Moda
27	58	64	65.35	75	17	89	-0.5664	0.2831	12.27	62

A amostra em estudo apresenta idades entre os 27 anos (mínima) e os 89 anos (máxima).

A mediana (50% dos dados) é 64 anos, indicando que metade das observações tem 64 anos ou menos e a outra metade tem 64 anos ou mais.

Ao considerar os quartis, observamos que 25% dos participantes têm menos de 58 anos, enquanto 75% têm menos de 75 anos. Isso destaca a diversidade de idades na amostra, com uma diferença de 17 anos entre o primeiro e o terceiro quartil. Entende-se assim que existe uma maior concentração de indivíduos com idades compreendidas entre os 64 anos e os 75 anos, com uma média de 65.35 anos.

A idade mais comum (valor que ocorre com maior frequência na amostra) é de 62 anos.

Analisando a forma da distribuição, verificamos uma ligeira assimetria para a esquerda, o que significa que há uma inclinação subtil para idades mais jovens. No entanto, a curtose próxima de zero sugere que a distribuição não é excessivamente achatada, indicando uma relativa uniformidade na dispersão das idades.

O desvio padrão de aproximadamente 12 anos destaca a variabilidade das idades em relação à média.

A amostra apresenta um afastamento da distribuição normal com alguma dispersão maior nos extremos etários.

A assimetria (skewness) é negativa, mas próxima de zero, embora com uma leve inclinação para a esquerda (enviesada à esquerda), indicando uma distribuição relativamente simétrica.

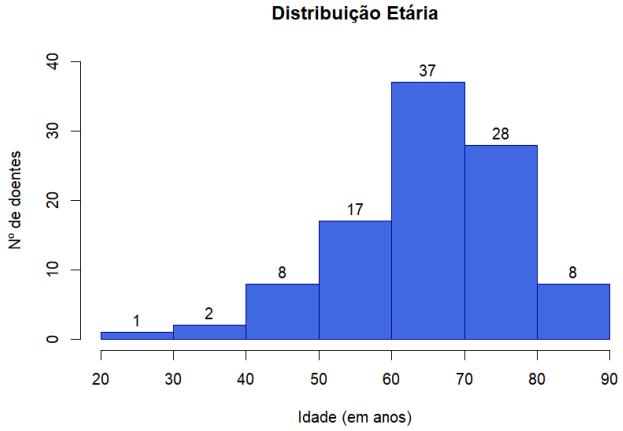


Gráfico 1 *- Histograma da variável "Idade"*

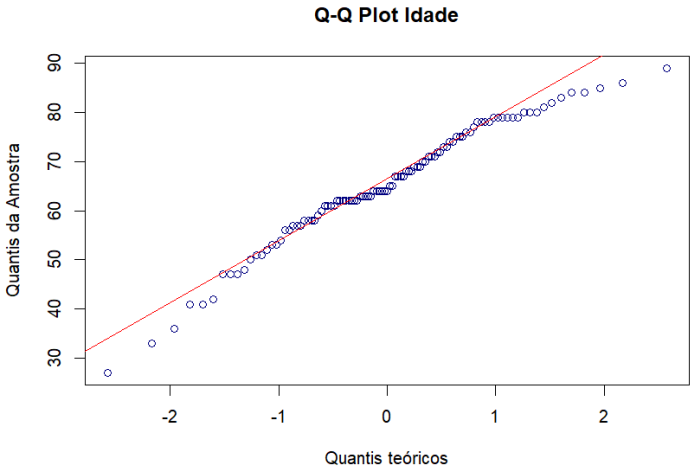


Gráfico 2 *- Q-Q Plot da variável "Idade"*

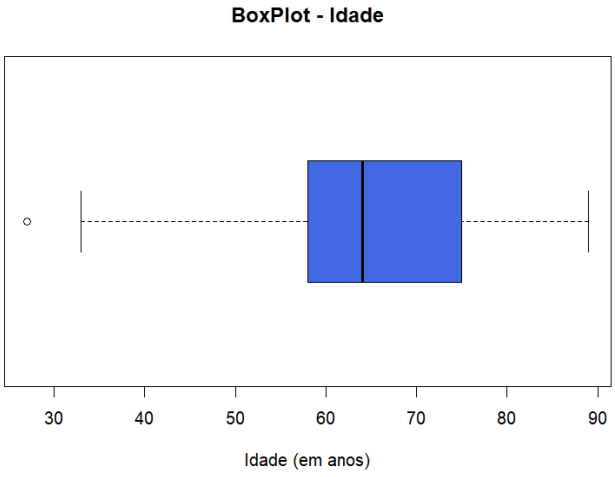


Gráfico 3 *- BoxPlot da variável "Idade"*

Para representar a variável, realizámos um histograma, um boxplot e um gráfico Q-Q.

No histograma (gráfico 1), verificamos uma distribuição relativamente simétrica em torno da média, com uma cauda mais longa à esquerda, indicando uma ligeira inclinação para idades mais jovens. As barras do histograma são mais altas nas idades mais comuns (em torno da moda de 62 anos), e a distribuição geral é razoavelmente uniforme, com algumas variações em torno da média.

No Q-Q plot (gráfico 2), de acordo também com os valores de curtose (próxima de zero) e a assimetria ligeiramente negativa, vemos uma linha aproximadamente reta, com uma inclinação subtil para a esquerda devido à assimetria. Verificamos que existem alguns pontos no centro da amostra que se alinham perto da linha diagonal, indicando que a distribuição de dados parece aproximadamente normal; contrapondo, existem ainda os pontos nos extremos que se afastam da linha normal, indicando desvio da normalidade – afastando-se em curva, sugerindo no seu todo uma distribuição diferente da normal.

No boxplot (gráfico 3), a caixa não é simétrica pelo que a linha que representa a mediana não se encontra no centro da mesma: está mais próxima de Q1 – verifica-se assim uma discreta assimetria negativa da distribuição. Dentro da caixa, entre Q1 e Q3, estão representados 50% dos dados analisados – a variação na metade central dos dados está na faixa dos 64 anos. É possível visualizar os valores extremos de idade e a presença de um outlier (27 anos), significativamente fora da distribuição normal.

Sexo

A variável **sexo** trata-se de uma variável qualitativa nominal, pelo que fará sentido percecionar como a amostra se divide em género. Realizámos uma tabela de frequências (absolutas e relativas), um gráfico de barras e um gráfico circular.

A estatística descritiva que fará sentido considerar nesta variável é a moda. Através da moda, identifica-se a categoria ou valor com maior frequência na variável em estudo. Neste sentido, o valor com maior frequência identifica-se como sendo o sexo Feminino, com uma frequência absoluta de 61 observações. Esta variável apresenta uma distribuição unimodal (ou seja, uma única moda).

***Tabela 3** - Tabela de Frequências - variável "Sexo"*

Sexo	n_i	f_i (%)
Masculino (1)	40	39,60
Feminino (2)	61	60,40
Total	101	100

Avaliando esta tabela retira-se que para a amostra em estudo, existem 61 indivíduos do sexo feminino e 40 indivíduos do sexo masculino, sendo 60,40% e 39,60% da amostra, respetivamente, conforme podemos também visualizar nos gráficos 4 e 5. A avaliação da tabela permite confirmar também o que se conclui com a avaliação da moda, apresentando uma frequência relativa entre categorias que difere em 20,8%.

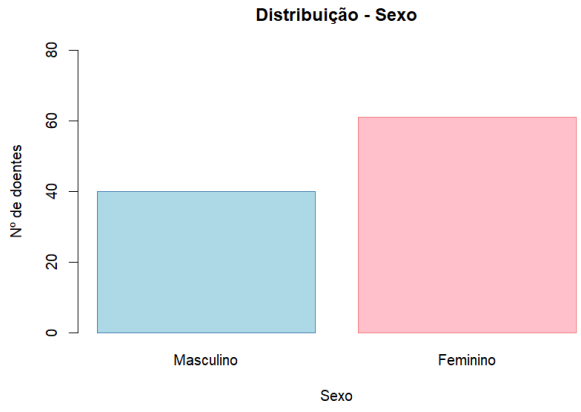


***Gráfico 4** - Gráfico de Barras - Sexo*

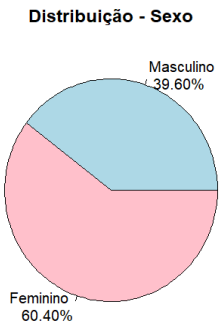


***Gráfico 5** - Gráfico Circular - Sexo*

Grupo_terap

A variável ***grupo_terap*** é qualitativa nominal, pelo que fará sentido percecionar como a amostra se divide por grupo terapêutico. Realizámos uma tabela de frequências (absolutas e relativas), um gráfico de barras e um gráfico circular.

Uma estatística descritiva que fará sentido considerar nesta variável é a moda. Através da moda, identifica-se a categoria ou valor com maior frequência na variável em estudo. Neste sentido, o valor com maior frequência identifica-se como sendo o grupo terapêutico C09: Agentes que atuam no sistema renina-angiotensina, com uma frequência absoluta de 71 observações. Esta variável apresenta, portanto, uma distribuição unimodal, com apenas uma única moda.

Tabela 4 *- Tabela de frequências - Grupo terapêutico*

Grupo terapêutico	n_i	f_i (%)
C02: Anti-hipertensivos	0	0
C03: Diuréticos	7	6,93
C07: Agentes bloqueadores beta	1	0,99
C08: Bloqueadores do canal de cálcio	4	3,96
C09: Agentes que atuam no sistema renina-angiotensina	71	70,30
2 ou mais	18	17,82
Total	101	100

Com a presente tabela de frequências consegue-se uma visualização rápida do número, percentagem ou proporção de doentes observados para cada uma das categorias em estudo, com o número de ocorrências (frequências absolutas) e com a proporção (frequência relativa) de observações que pertencem a cada categoria (grupo terapêutico).

Desta tabela consegue-se confirmar que de facto a moda (valor com mais ocorrências) corresponde ao grupo terapêutico C09: Agentes que atuam no sistema renina-angiotensina, com uma frequência absoluta de 71 e uma percentagem face ao total de ocorrências de 70,30%. De seguida, aparece a categoria "2 ou mais" (grupos terapêuticos), com 18 observações, correspondendo a um valor relativo de 17,82% face ao total de todas as categorias.

Na amostra em estudo, não se observa o grupo terapêutico C02: Anti-hipertensivos. O grupo terapêutico com menor frequência corresponde ao grupo C07: agentes bloqueadores beta com apenas 1 observação, correspondente a 0,99% do total. As restantes categorias distribuem-se de forma mais ou menos homogénea em termos de frequência, com uma amplitude de cerca de 3% de ocorrências entre elas.

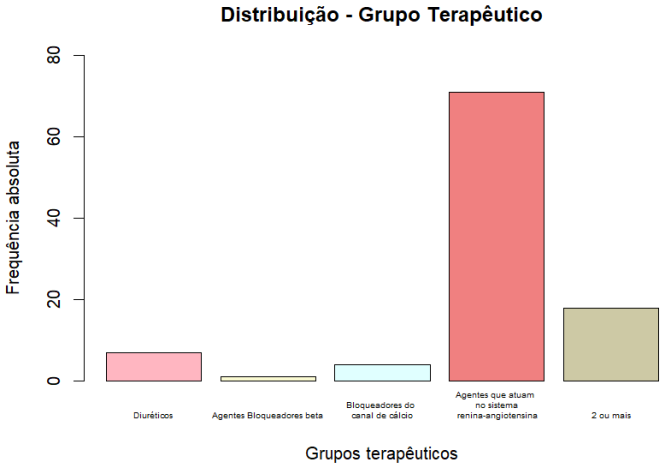


Gráfico 6 *- Gráfico de Barras da variável "Grupo Terapêutico"*

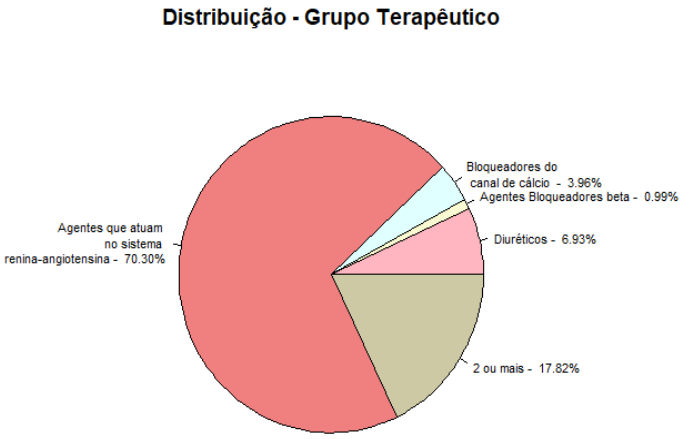


Gráfico 7 *- Gráfico circular da variável "Grupo Terapêutico"*

Através dos gráficos acima (gráficos 6 e 7) podemos extrair as mesmas conclusões anteriormente referidas, de uma forma mais visual. Note-se que graficamente se observa uma frequência consideravelmente superior da categoria Grupo terapêutico C09: Agentes que atuam no sistema renina-angiotensina, correspondendo à barra mais elevada do conjunto assim como à fatia do gráfico circular, apresentando uma proporção muito superior às restantes.

Nr_receita

A variável ***nr_receita*** é qualitativa ordinal, representando o número de medicamentos na primeira receita apresentando os possíveis valores: 1=um ou 2=dois ou mais, pelo que fará sentido percecionar como a amostra se divide. Realizámos uma tabela de frequências (absolutas e relativas), um gráfico de barras e um gráfico circular.

A estatística que fará também sentido considerar nesta variável é a moda. Através da moda, identifica-se a categoria ou valor com maior frequência na variável em estudo. Neste sentido, interpreta-se que o valor com maior frequência para o número de medicamentos na primeira receita identifica-se como sendo o um (1), com uma frequência de 83 observações. Esta variável apresenta, portanto também, uma distribuição unimodal, com apenas uma única moda.

Tabela 5 - Tabela de Frequências - Número de medicamentos na primeira receita

Nr_receita	n_i	f_i (%)
Um (1)	83	82,18
Dois ou mais (2)	18	17,82
Total	101	100

Pela análise desta tabela de frequências conclui-se que a moda toma o valor *um (1)*, com uma frequência absoluta de 83 observações e frequência relativa de 82,18%. Em contrapartida o valor *dois ou mais (2)* apresenta respetivamente 18 observações correspondendo a 17,82% do total de registos. Estas conclusões são também visíveis nos gráficos abaixo (gráficos 8 e 9): existe uma clara predominância para a primeira receita dos doentes apresentarem apenas um medicamento AHT.

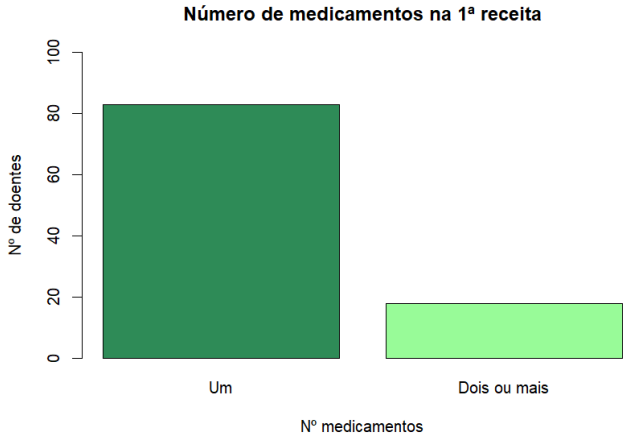


Gráfico 8 - Gráfico de Barras da variável "Nr_receita"

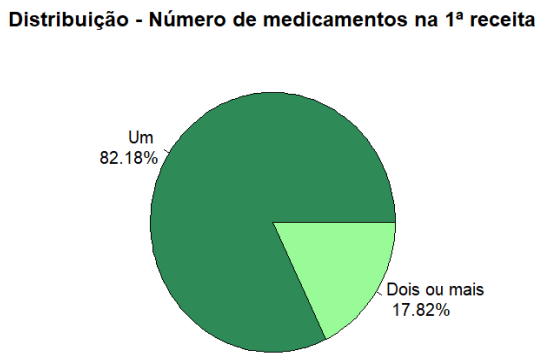


Gráfico 9 - Gráfico circular da variável "Nr_receita"

3. Recorrendo a gráficos, compare o tempo até à primeira aquisição (tempo_inicio) por grupo terapêutico (grupo_terap).

Tabela 6 *- Estatísticas "Tempo até à primeira aquisição" por "Grupo terapêutico"*

Grupo terapêutico	n_i	Min	1ºQ	Mediana	$\bar{x}$	3ºQ	Max	NA
C02: Anti-hipertensivos	0	0	0	0	0	0	0	0
C03: Diuréticos	7	2	6	17	12.6	18	20	2
C07: agentes bloqueadores beta	1	41	41	41	41	41	41	41
C08: bloqueadores do canal de cálcio	4	21	23	25	43.33	54.50	84	1
C09: Agentes que atuam no sistema renina-angiotensina	71	0	10	18	22.48	28	106	6
2 ou mais	18	0	6	20	25.47	41	80	3
Total	101							

Para comparar o tempo até à primeira aquisição por grupo terapêutico, importa uma representação gráfica como o boxplot (gráfico 10), onde podemos retirar algumas conclusões de forma mais visual, espelhando alguns dos dados da tabela 6.

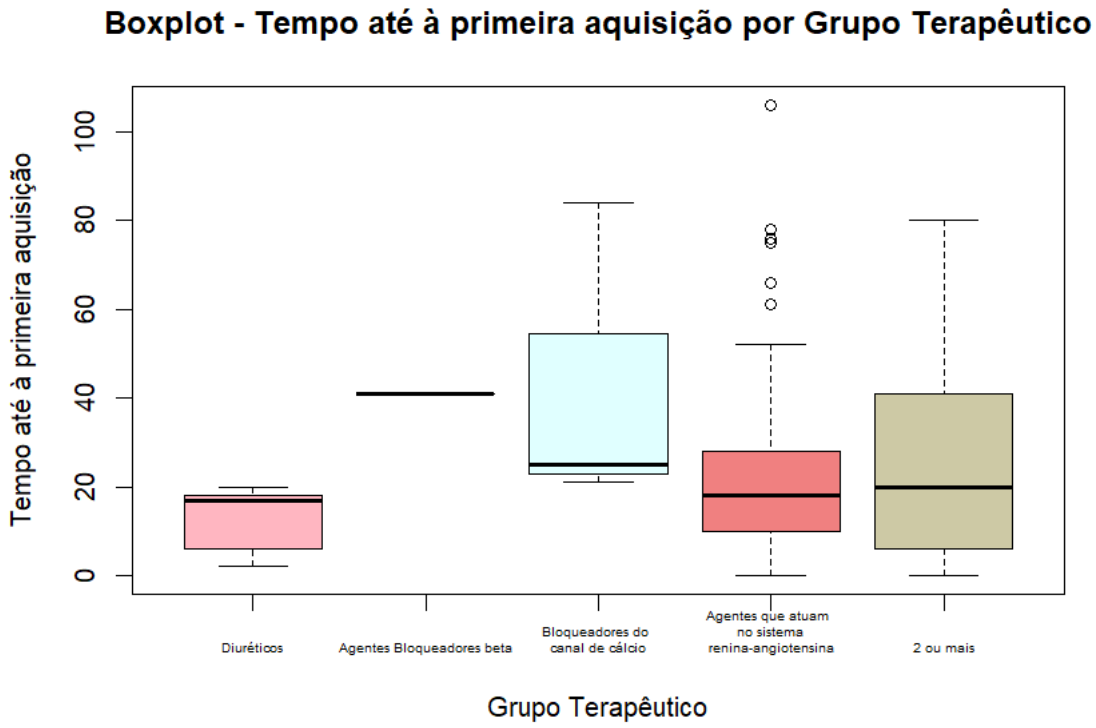


Gráfico 10 *-Boxplot "Tempo até à primeira aquisição" por "Grupo terapêutico"*

- **C02: Anti-hipertensivos:** Como não há dados na amostra para este grupo, o gráfico boxplot não é gerado.
- **C03: Diuréticos:** O tempo até a primeira aquisição para diuréticos varia amplamente, com uma mediana de 17 dias e uma quantidade significativa de variação entre os indivíduos.
- **C07: Agentes bloqueadores beta:** Há apenas um dado neste grupo, resultando num boxplot simples que mostra que a primeira aquisição ocorreu aos 41 dias.
- **C08: Bloqueadores do canal de cálcio:** Este grupo apresenta uma grande dispersão nos tempos até a primeira aquisição, com uma mediana de 25 dias e uma faixa que se estende até 84 dias.
- **C09: Agentes que atuam no sistema renina-angiotensina:** Apresenta uma mediana de 18 dias e uma distribuição mais concentrada em comparação com os grupos anteriores, mas com uma variação considerável – apresenta vários valores outliers, ao contrário dos restantes grupos.
- **2 ou mais:** Este grupo também mostra uma ampla variação nos tempos até a primeira aquisição, com uma faixa que se estende até 80 dias e com uma mediana de 20 dias – a mediana apresenta um valor aproximado ao grupo anterior, mas tem uma distribuição consideravelmente diferente.

Consideramos também interessante a representação gráfica do tempo médio até à primeira aquisição por grupo terapêutico (gráfico 11), concluindo que, pela quantidade de observações de cada grupo, nomeadamente do grupo Agentes bloqueadores beta onde apenas se verifica uma observação na amostra, existe um afastamento considerável em relação à mediana, não sendo esta a melhor medida a aplicar à nossa amostra numa ótica de comparação por não refletir a realidade da população.

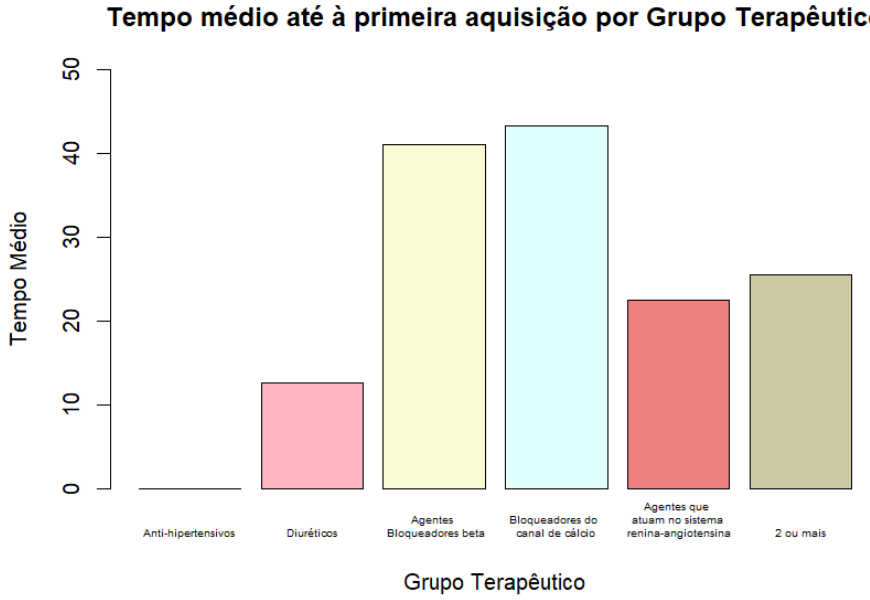


***Gráfico 11**- Tempo médio até à primeira aquisição por "Grupo terapêutico"*

- **C02: Anti-hipertensivos:** Sem observações (0) na amostra.
- **C03: Diuréticos:** A média é de aproximadamente 12.6 dias, enquanto a mediana é de 17 dias. Isso sugere que há uma leve assimetria à esquerda na distribuição dos dados, com uma pequena quantidade de valores menores a influenciar a média para baixo em relação à mediana.
- **C07: Agentes bloqueadores beta:** Tanto a média quanto a mediana são de 41 dias por apenas se registar uma observação na amostra.
- **C08: Bloqueadores do canal de cálcio:** A média é de aproximadamente 43.33 dias, enquanto a mediana é de 25 dias. Isso sugere que pode haver uma assimetria à direita na distribuição, com alguns valores extremamente altos a puxar a média para cima em relação à mediana.
- **C09: Agentes que atuam no sistema renina-angiotensina:** A média é de aproximadamente 22.48 dias, enquanto a mediana é de 18 dias. Isso indica uma ligeira assimetria à direita na distribuição, com alguns valores mais altos aumentando a média em relação à mediana.
- **2 ou mais:** A média é de aproximadamente 25.47 dias, enquanto a mediana é de 20 dias. Isso sugere que pode haver uma assimetria à direita na distribuição, com alguns valores mais altos a puxar a média para cima em relação à mediana.

4. Obtenha uma representação gráfica que permita comparar a compra de medicamentos genéricos ou marca (marca_generico) por sexo e interprete-o.

***Tabela 7**- Frequência absoluta da compra de medicamentos Genéricos ou Marca por Sexo*

Sexo/ Generico_marca	Masculino	Feminino	
Genérico	12	25	
Marca	21	25	
NA	7	11	
Total	40	61	101

Para melhor analisar e visualizar a distribuição dos dados, recorremos a um gráfico de barras (gráfico 12) que diferencia a amostra por sexo, interpretando a quantidade de medicamentos consumidos por genérico ou marca para cada um dos mesmos.

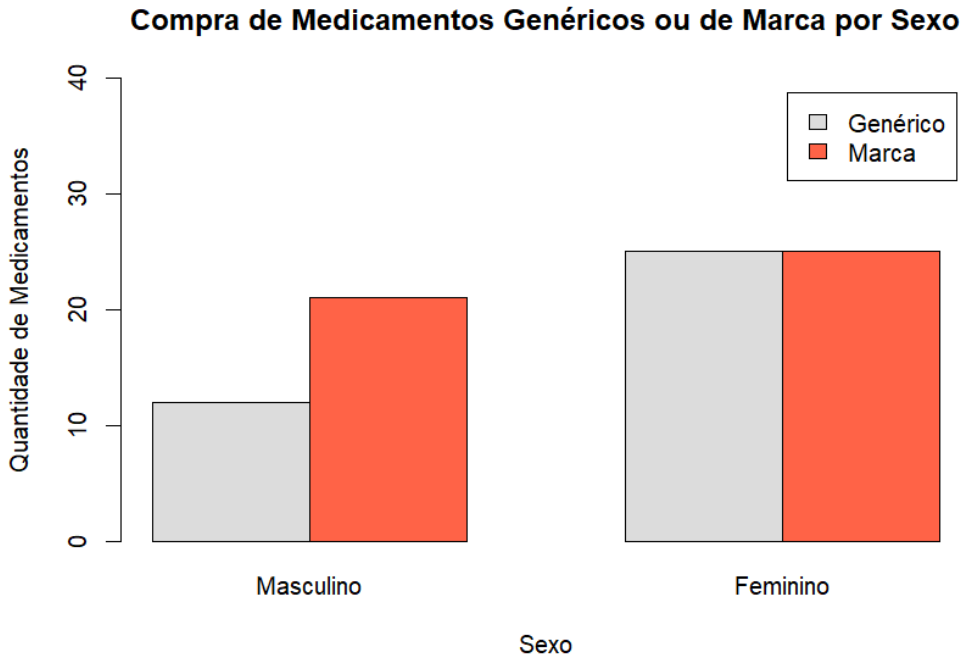


***Gráfico 12** - Compra de Medicamentos genéricos ou de marca por sexo*

Observa-se que o número total de indivíduos do sexo feminino que compraram medicamentos é superior ao número de indivíduos do sexo masculino (61 contra 40, respetivamente).

Essa diferença no total de compradores deve-se a maior prevalência do que os homens na amostra, conforme analisado anteriormente.

Analisando a preferência por Medicamentos Genéricos e de Marca note-se que, entre os doentes do sexo masculino, houve 12 compras de medicamentos genéricos, enquanto 21 compraram medicamentos de marca.

No sexo feminino, não se observou nenhuma preferência por medicamentos genéricos nem de marca por igualdade de número de compras de cada tipo, ambas com total de 25 compras.

De salientar que houve doentes que não informaram sua escolha ou cujos dados não foram registados (NA).

Tanto entre os homens quanto entre as mulheres, houve um número representativo de casos em que os indivíduos não informaram se compraram medicamentos genéricos ou de marca.

Para o sexo feminino, o número de valores omissos foi ligeiramente maior em comparação contando com cerca de 18,3% em comparação com o grupo masculino com 17,5% (11 e 7 doentes, respetivamente).

A partir destes dados, podemos inferir que enquanto os homens tendem a comprar mais medicamentos de marca, as mulheres não demostram preferência.

No entanto, a proporção de valores omissos pode considerar-se relevante em ambos os sexos, sugerindo que uma parte considerável dos consumidores pode não ter uma preferência clara entre medicamentos genéricos e de marca, ou que houve falta de registo dessas informações.

Esta interpretação dos resultados destaca padrões de compra de medicamentos entre os sexos e destaca a importância de considerar não apenas as preferências individuais, mas também a importância da precisão e a integralidade dos registos de compra para uma análise mais completa.

5. Construa um diagrama de dispersão entre o poder de compra (Poder-compra) e o custo inicial (custo_inicial) e conclua. Complemente com uma estatística que considere adequada.

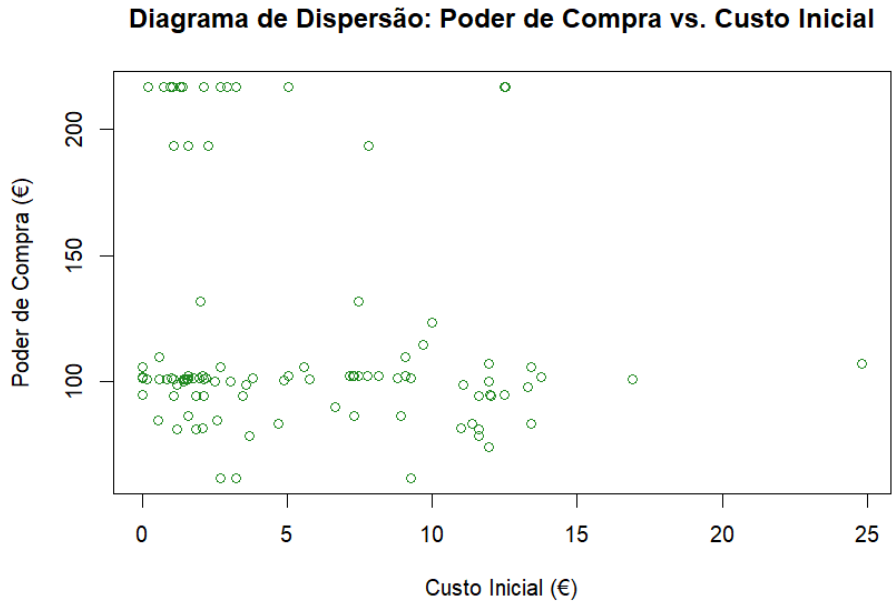


***Gráfico 13**- Diagrama de dispersão - Poder de compra VS. Custo Inicial*

No diagrama de dispersão (gráfico 13), observamos que não há uma tendência clara nos dados. Os pontos estão espalhados de forma irregular e não parece haver uma relação linear evidente entre o custo inicial e o poder de compra.

Para decidir que tipo de coeficiente de correlação devemos aplicar (sendo duas variáveis quantitativas), é necessário avaliar se ambas têm distribuição normal. Para isso, realizamos para ambas um boxplot (gráficos 14 e 15), um gráfico Q-Q Plot (gráficos 16 e 17), calculámos a assimetria e curtose, finalizando com o teste Shapiro-Wilk.

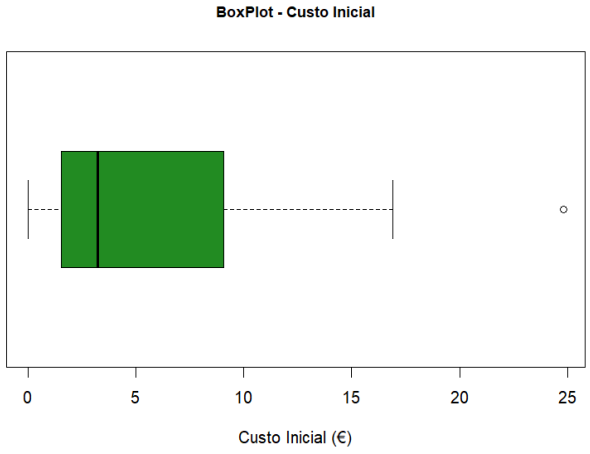


***Gráfico 14**- Boxplot Custo Inicial*

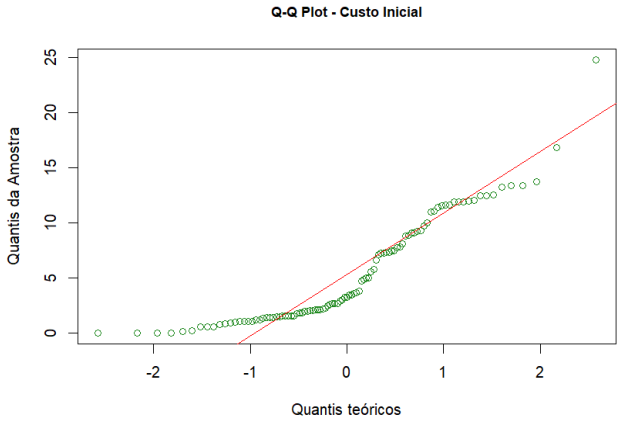


***Gráfico 15**- Q-Q plot Custo Inicial*

1. **Skewness (Assimetria):** O valor de skewness (assimetria) para a variável "custo inicial" é de aproximadamente 1.0670. Uma assimetria positiva indica que a cauda direita da distribuição é mais longa do que a cauda esquerda, sugerindo uma distribuição assimétrica para a direita, conforme também espelhado no boxplot (gráfico 14), agravado com a presença de um valor outlier.
2. **Kurtosis (Curtose):** O valor de kurtosis (curtose) para a variável "custo inicial" é de aproximadamente 1.0622. A curtose mede a cauda de uma distribuição. Um valor de curtose maior que zero indica que a distribuição tem caudas mais pesadas do que uma distribuição normal, o que significa que há mais dados nos extremos do que seria de esperar numa distribuição normal.
3. **Teste de Normalidade de Shapiro-Wilk:** O teste de normalidade de Shapiro-Wilk é um teste estatístico utilizado para determinar se uma amostra de dados segue uma distribuição normal. No caso da variável "custo inicial", o valor-p é muito pequeno (Valor p <.001), o que sugere que há evidências significativas para rejeitar a hipótese nula de que os dados seguem uma distribuição normal, algo confirmado também pelos gráficos – no Q-Q plot (gráfico 15), verificamos que existem alguns pontos no centro da amostra que se alinham perto da linha diagonal mas existem ainda os pontos nos extremos que se afastam da linha normal, indicando desvio da normalidade – afastando-se em curva, sugerindo no seu todo uma distribuição diferente da normal. Portanto, a distribuição dos dados da variável “custo inicial” não é normal.

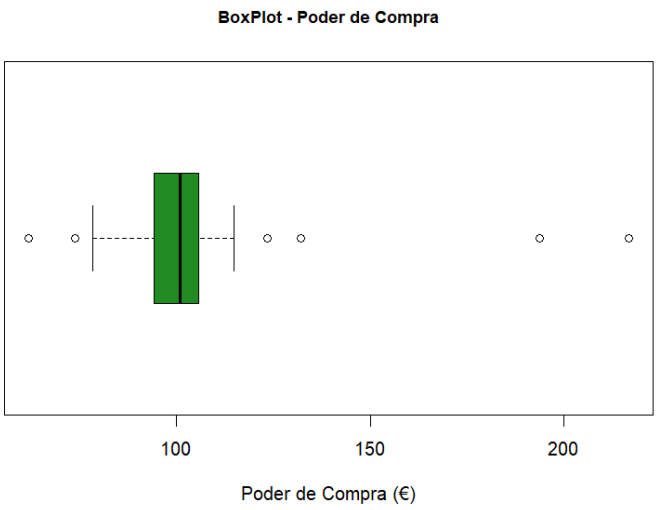


Gráfico 16- Boxplot Poder de Compra

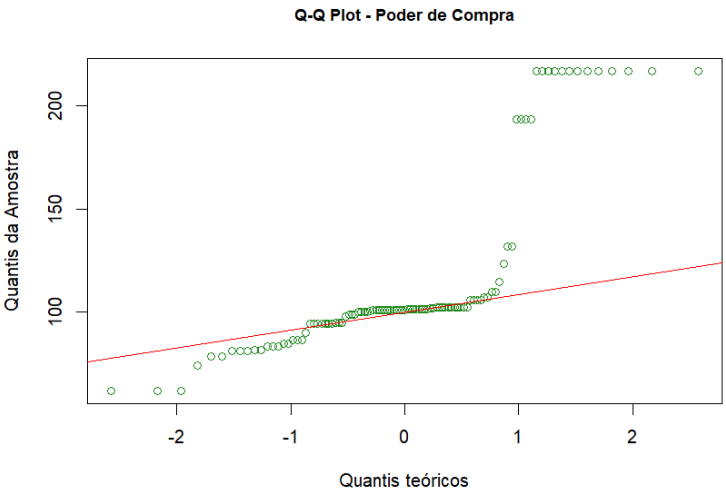


Gráfico 17- Q-Q Plot Poder de Compra

1. **Skewness (Assimetria):** O valor de skewness (assimetria) para a variável "poder de compra" é de aproximadamente 1.5482. Uma assimetria positiva indica que a cauda direita da distribuição é mais longa do que a cauda esquerda, o que sugere uma distribuição assimétrica para a direita.
2. **Kurtosis (Curtose):** O valor de kurtosis (curtose) para a variável "poder de compra" é de aproximadamente 0.8899. A curtose mede a cauda de uma distribuição. Um valor de curtose maior que zero indica que a distribuição tem caudas mais pesadas do que uma distribuição normal, o que significa que há mais dados nos extremos do que seria de esperar numa distribuição normal.
3. **Teste de Normalidade de Shapiro-Wilk:** O teste de normalidade de Shapiro-Wilk é um teste estatístico utilizado para determinar se uma amostra de dados segue uma distribuição normal. No caso da variável "poder compra", o valor-p é muito pequeno (Valor p <.001) o que sugere que há evidências significativas para rejeitar a hipótese nula de que os dados seguem uma distribuição normal, conforme também possível inferir através dos gráficos à semelhança da variável “custo inicial”. Portanto, a distribuição dos dados da variável "poder de compra" não é normal.

Sabendo agora que nenhuma das variáveis tem uma distribuição normal, devemos aplicar o coeficiente de correlação de Spearman.

A correlação é uma medida estatística que descreve a relação entre duas variáveis quantitativas. É calculada para determinar a direção e a força dessa relação. Tanto o custo inicial quanto o poder de compra são variáveis numéricas que representam medidas quantitativas. Isso torna a correlação uma ferramenta estatística apropriada para analisar a relação entre ambas, sendo que a correlação permite-nos quantificar essa relação e determinar se ela é positiva, negativa ou nula.

Correlação entre Custo Inicial e Poder de Compra (Spearman): -0.1218062

Podemos concluir que há uma correlação negativa fraca (pouca ou nenhuma associação) entre as duas variáveis.

Isso significa que, de forma geral, à medida que o custo inicial aumenta, o poder de compra tende a diminuir, embora essa relação não seja muito forte. Sugere uma relação ligeiramente inversa entre o custo inicial das receitas e o poder de compra, mas também indica que outros fatores podem estar a influenciar o poder de compra além do custo inicial das receitas.

6. Verifique se existe associação entre a variável "inicio" e o código ICPC.

Para respondermos a esta questão temos de recorrer a estatística bivariada, que tem como objetivo verificar o nível de relação entre duas variáveis. Para tal, é necessário caracterizar as variáveis a analisar de forma a aplicar os métodos estatísticos adequados.

Variáveis a avaliar – Início, ICPC: código da hipertensão – variáveis qualitativas nominais, ambas dicotómicas (ou seja, apresentam apenas duas classes em cada categoria).

Os métodos estatísticos adequados serão:

- Tabela de contingência
- Coeficiente de associação: Coeficiente Phi.

Para analisarmos a associação entre estas 2 variáveis, é necessário construir uma tabela de contingência de frequência relativas (em percentagem) através da função *prop.table* (tabela.

A tabela de contingência auxilia a organização e visualização dos dados em relação as variáveis qualitativas, como é o caso, permitindo uma melhor interpretação. Este passo é necessário para o cálculo do coeficiente de associação, a partir da tabela de contingência.

Sendo duas variáveis qualitativas nominais, ambas dicotómicas, a existência de associação pode ser verificada através do coeficiente Phi.

Tabela 8- *Frequências relativas (início e ICPC)*

	K86 – hipertensão sem complicações (%)	K87 – hipertensão com complicações (%)	Total (%)
Não inicia tratamento	0	0	0
Inicia tratamento	93.51	6.49	100
Total	93.51	6.49	100

Através da tabela de contingência (tabela 8), verificamos que, na nossa amostra, não se observam registos sobre os doentes que não iniciaram tratamento, não se verificando valores de hipertensão sem complicações nem hipertensão com complicações. Não existindo dados que permitam verificar a associação na amostra em estudo, não conseguimos verificar associação entre as variáveis.

7. Analise as variáveis descontinua, persist_6m, persist_24m, descontinua2, recorrendo a estatísticas e/ou gráficos e conclua sobre a adesão à terapêutica dos pacientes.

Para as variáveis em estudo, pretendemos verificar e comparar a adesão terapêutica dos doentes da nossa amostra. Ambas as variáveis apresentam valores omissos, pelo que o tratamento dos dados terá esse facto em conta.

***Tabela 9**- Frequências - descontinua, persist_6m, persis_24m, descontinua2*

Variáveis		n_i	f_i (%)	Moda
Descontinua	Sim	9	10.11	Não
	Não	80	89.89	
Persist_6m	Sim	78	87.64	Sim
	Não	11	12.36	
Persist_24m	Sim	69	77.53	Sim
	Não	20	22.47	
Descontinua2	Sim	20	22.47	Não
	Não	69	77.53	
Total indivíduos		89	100%	

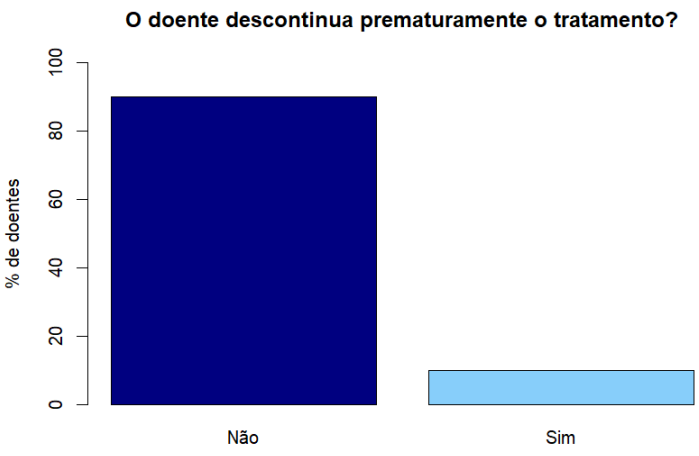


***Gráfico 18** - Doentes que descontinuam prematuramente o tratamento*

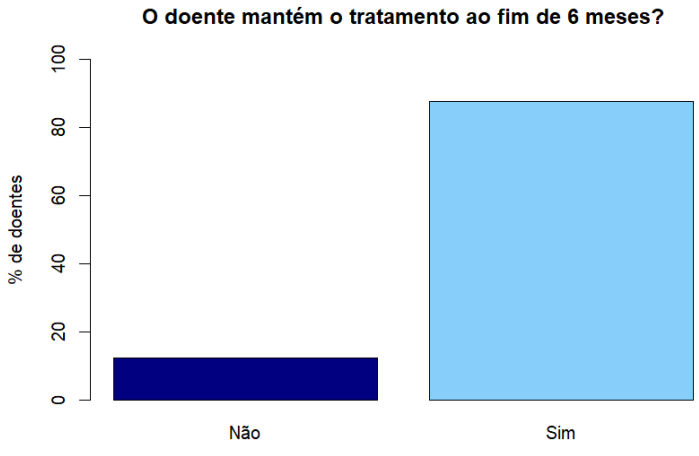
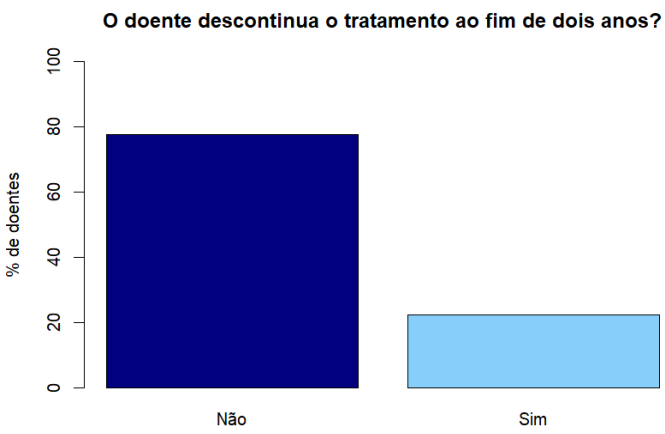


***Gráfico 19** - Doentes que mantém tratamento ao fim de 6 meses*



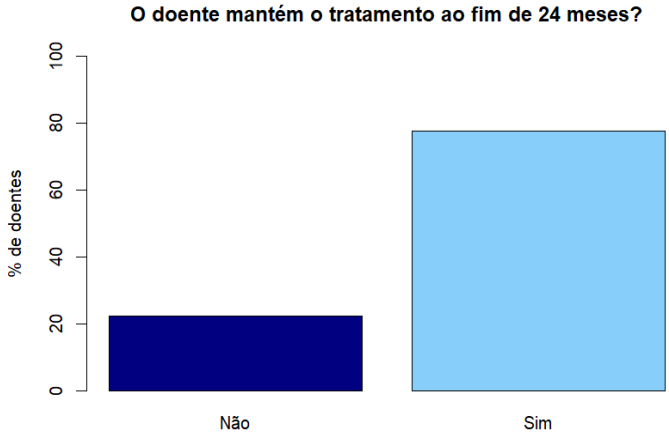


***Gráfico 20**- Doentes que descontinuam o tratamento ao fim de dois anos* ***Gráfico 21** - Doentes que mantém tratamento ao fim de 24 meses*

Através da análise das variáveis "Persist_24m" e "Descontinua2" e do que cada uma representa, conseguimos perceber que a janela temporal a cada uma associada é a mesma, ou seja, 2 anos / 24 meses. Um doente que descontinue o tratamento ao fim de dois anos, é simultaneamente um doente que não persiste no tratamento ao fim de 24 meses; da mesma forma que um doente que persista no tratamento ao fim de 24 meses, é simultaneamente um doente que não descontinua o tratamento ao fim de 2 anos.

Com os resultados obtidos, podemos concluir que a maioria dos pacientes não descontinuaram precocemente a terapia de hipertensão arterial e que este padrão se mantém próximo ao fim de dois anos, persistindo no tratamento. A maioria dos pacientes (77,53%) permaneceu na terapia por 24 meses e não descontinuou o tratamento.

As variáveis "Descontinua" e "Persist_6m" apresentam também o mesmo padrão das referidas anteriormente, ainda que pelos resultados apresentados, não haja uma correspondência direta entre elas, nomeadamente por não termos informação complementar sobre o tempo associado à precocidade.

A representação gráfica de cada variável (gráficos 18 a 21) mostra visualmente e corrobora a análise realizada acima, bem como a moda.

Em suma, os resultados sugerem uma adesão relativamente forte à terapia de hipertensão arterial entre os pacientes analisados, demonstrando o seguimento do tratamento de uma forma contínua no tempo.

De considerar conforme analisado anteriormente que, para as variáveis em estudo na adesão à terapêutica, não temos representação de 12 pacientes.

8. Obtenha uma estimativa pontual para o valor médio do poder de compra (Poder_compra) e uma estimativa intervalar com uma confiança de 99% apenas para os indivíduos que aviam dois ou mais medicamentos.

Estimativa pontual:

Para calcular esta estimativa pontual, pretendemos fazer a média do poder de compra apenas para o subconjunto de indivíduos que aviam dois ou mais medicamentos, e a partir daí verificar o valor médio do poder de compra para essa subamostra.

A estimativa pontual para o valor médio do poder de compra para os indivíduos que aviam dois ou mais medicamentos é de 116,17(€).

Estimativa intervalar a 99%:

De forma a eleger o tipo de teste que deve ser aplicado para realizar uma estimativa intervalar a 99%, é necessário avaliar diversos parâmetros entre eles a dimensão da amostra e o tipo de distribuição.

Verificar critérios:

- X v.a = poder de compra para doentes que aviam dois ou mais medicamentos
- Variável X tem uma distribuição desconhecida
- Dimensão da amostra *(n)*? *n* = 18 – distribuição tem de ser normal para se poder fazer a estimativa.
- σ desconhecido

Verificar normalidade da distribuição:

***Tabela 10**- Estatísticas - Poder de compra para doentes que aviam dois ou mais medicamentos*

X v.a	Min	1ºQ	Mediana	$\bar{x}$	3ºQ	Max
Poder de compra para doentes que aviam dois ou mais medicamentos	61.72	94.13	100.85	115.59	104.90	216.88

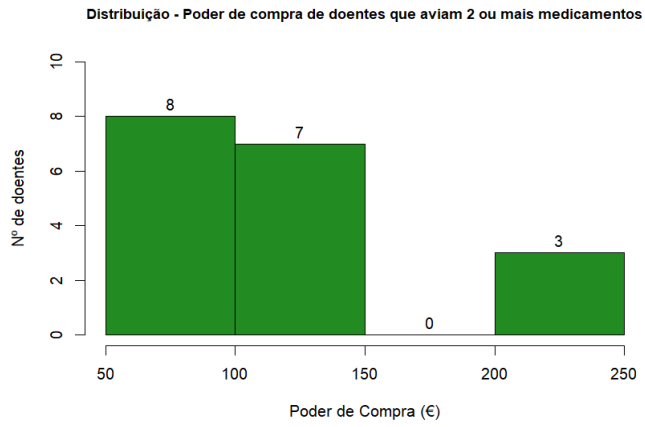


***Gráfico 22** - Histograma Poder de compra de doentes que aviam 2 ou mais medicamentos*

Pela análise do histograma (gráfico 22), a distribuição da variável em causa não parece simétrica, com uma assimetria positiva. A cauda da distribuição estende-se mais para a direita, o que significa que a maioria dos valores está concentrada na parte esquerda da distribuição.

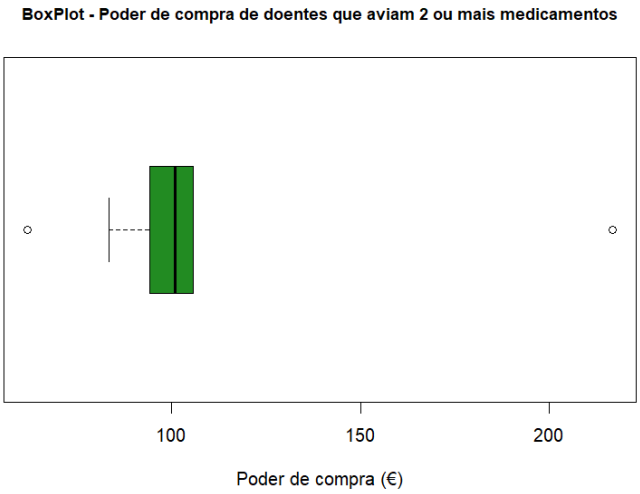


Gráfico 23 *- Boxplot Poder de compra de doentes que aviam 2 ou mais medicamentos*

No boxplot (gráfico 23), verificamos uma distribuição mais acentuada dos valores à esquerda. Verifica-se também a existência de outliers, com valor 61.72€ e 216.88€ de poder de compra. Observa-se também que 75% destes doentes apresentam um poder de compra abaixo de 104.90€. Não se verificam valores distribuídos entre 150€ e 200€, conforme observado no histograma.

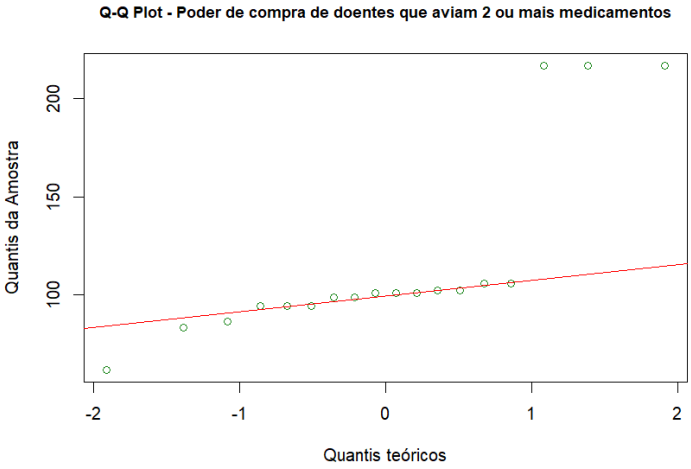


Gráfico 24 *- Q-Q Plot Poder de compra de doentes que aviam 2 ou mais medicamentos*

Avaliando o gráfico Q-Q Plot (gráfico 24), observam-se muitos valores sobre a linha de normalidade, mas algumas observações encontram-se bastante desviadas. Nesse sentido, vamos testar com o Shapiro-Wilk.

1. **Skewness (Assimetria):** O valor de skewness (assimetria) é de aproximadamente 1.455 - indica uma assimetria positiva na distribuição dos dados - significa que a cauda direita da distribuição é mais longa do que a cauda esquerda, podendo afirmar também que os valores estão mais concentrados à esquerda da média e que há alguns valores extremamente altos que estendem a distribuição para a direita.
2. **Kurtosis (Curtose):** O valor de kurtosis (curtose) é de aproximadamente 0.502. A curtose mede a cauda de uma distribuição. Um valor de curtose maior que zero indica que a distribuição tem caudas mais pesadas do que uma distribuição normal, o que significa que há mais dados nos extremos do que seria de esperar numa distribuição normal.
3. **Teste de Normalidade de Shapiro-Wilk:** O teste de normalidade de Shapiro-Wilk é um teste estatístico utilizado para determinar se uma amostra de dados segue uma

distribuição normal. No caso da variável "poder de compra" de doentes que aviam 2 ou mais medicamentos, o valor-p é muito pequeno (Valor p <.001) o que sugere que há evidências significativas para rejeitar a hipótese nula de que os dados seguem uma distribuição normal, conforme também possível inferir através dos gráficos à semelhança da variável "custo inicial". Portanto, a distribuição dos dados da variável "poder de compra" desta subamostra não é normal.

Não se verificando uma distribuição normal, a abordagem deverá ser não paramétrica, não abordada no conteúdo programático da unidade curricular.

9. Compare a estimativa por intervalo de confiança a 95% para o mesmo parâmetro da alínea anterior e conclua.

Conforme verificámos na questão anterior, não estamos em condições para calcular a estimativa intervalar por meio da abordagem não paramétrica necessária. Neste sentido, não poderemos dar resposta ao exercício. Ainda assim, pelos conteúdos teóricos lecionados acerca dos intervalos de confiança, entendemos que o pretendido é comparar a estimativa com intervalos de confiança diferentes.

O intervalo de confiança é uma faixa de valores que é usada para estimar um parâmetro desconhecido de uma população com base numa amostra dos dados, fornecendo uma medida da precisão da estimativa e indicando a margem de erro esperada.

A diferença entre um intervalo de confiança de 99% e um intervalo de confiança de 95% reside na sua amplitude e na sua confiança estatística.

Enquanto um intervalo de confiança de 99% apresenta uma margem de erro de 1%, o intervalo de 95% dá-nos uma margem de erro de 5%. A margem de erro é uma medida estatística que quantifica o nível de incerteza associado a uma estimativa feita a partir de uma amostra em relação à verdadeira característica da população. Quanto menor a margem de erro, maior a precisão da estimativa sendo mais provável que a verdadeira média populacional esteja próxima à estimativa da amostra, ou seja, para um intervalo de confiança de 95%, há uma probabilidade de 95% de que o intervalo de confiança calculado a partir da amostra capture o verdadeiro parâmetro populacional.

Em oposição, uma margem de erro maior, indica maior incerteza e precisão sendo e a verdadeira média populacional pode variar mais amplamente. Um intervalo de confiança de 99% significa que, se se repetir a amostragem muitas vezes e calcular o intervalo de confiança para cada amostra, em 99% das vezes, o intervalo conterá o verdadeiro valor do parâmetro populacional. Portanto, um intervalo de confiança de 99% é mais amplo do que um intervalo de confiança de 95% porque se está mais confiante de que o intervalo capturará o verdadeiro valor do parâmetro, o que resulta numa margem de erro maior.

Em suma, um intervalo de confiança de 99% oferece uma confiança maior na inclusão do verdadeiro parâmetro populacional, mas a custo de um intervalo mais amplo.

Neste sentido, seria esperado essa observarmos esta diferença de resultados obtidos a partir destes dois exercícios, na medida em que para o intervalo de confiança a 99% verificaríamos um intervalo de valores mais amplo, contudo mais próximo do verdadeiro valor do parâmetro que se pretende.

10. Considerando apenas os indivíduos com idade inferior ou igual a 45 anos, pode afirmar-se que o valor médio dos valores obtidos para o custo inicial (Custo_inicial) é significativamente superior a 3?

Para verificar esta afirmação, aplicamos um teste de hipóteses.

Um teste de hipóteses é uma aplicação estatística que permite testar uma afirmação sobre um parâmetro ou característica de uma população com base em informações retiradas de uma amostra aleatória dessa população, como é o caso.

Envolve a formulação de duas hipóteses: a hipótese nula (H0) e a hipótese alternativa (H1). O objetivo é testar a hipótese que vai ser testada H0 contra a hipótese alternativa H1. A estatística de teste é utilizada para tomar uma decisão relativamente à hipótese nula (rejeitar ou não rejeitar).

Para a nossa amostra em estudo (n), pretende-se apenas testar esta afirmação para os indivíduos com idade inferior ou igual a 45 anos, pelo que dividimos a amostra pelos doentes com idade igual ou inferior a 45 anos (< 46 anos). Este subconjunto conta com uma amostra de 6 doentes.

Caracterização da variável (X v.a) – Representa o valor médio dos valores obtidos para o custo inicial em indivíduos com idade inferior ou igual a 45 anos.

Nível de significância – 5% ($\alpha = 0,05$) (conforme o enunciado)

Formulação das hipóteses:

Hipótese nula (H0): O valor médio dos custos iniciais para os indivíduos com idade inferior ou igual a 45 anos é menor ou igual a 3 (ou seja, H0: $\mu \leq 3$).

Hipótese alternativa (H1): O valor médio dos custos iniciais para os indivíduos com idade inferior ou igual a 45 anos é significativamente superior a 3 (ou seja, H1: $\mu > 3$).

Com o objetivo de definir as características em estudo, como a amostra, a média pontual e a normalidade da amostra bem como verificar o teste de hipóteses adequado, realizamos o estudo recorrendo ao R Studio, obtendo os seguintes.

Cálculos auxiliares:

N=6

$\bar{x}$ pontual = 8.33

Uma vez que n < 30, é necessário verificar se a variável em estudo é normal, através do gráfico Q-Q Plot (gráfico 25), confirmando posteriormente também com o teste de normalidade Shapiro-Wilk.

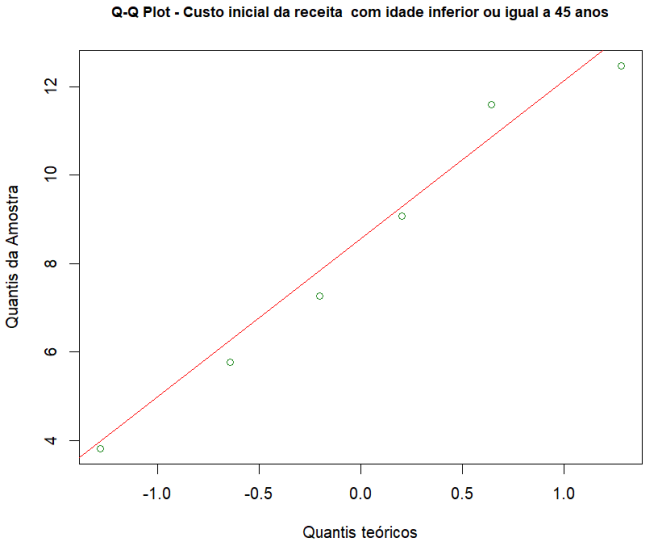


***Gráfico 25** -Q-Q Plot Poder de compra de doentes que aviam 2 ou mais medicamentos*

A partir do gráfico Q-Q plot (gráfico 25) sugere-se uma distribuição relativamente normal com muitos registos perto da linha de normalidade. Para confirmar esta sugestão, recorremos ao teste de Shapiro-Wilk.

O teste de Shapiro-Wilk é uma ferramenta estatística utilizada para verificar se uma amostra de dados segue uma distribuição normal. Este teste específico testa a hipótese nula de que a amostra tem distribuição normal. O teste resulta em um valor-p, que é utilizado para determinar a significância estatística do teste. Se o valor-p for menor que o nível de significância previamente escolhido, a hipótese nula é rejeitada e há evidência de que os dados não seguem uma distribuição normal. Por outro lado, se o valor p for significativamente maior que o nível de significância, a hipótese nula não pode ser rejeitada indicando que os dados podem ser considerados normalmente distribuídos.

Teste de Shapiro-wilk: valor p = 0.818 (> α)

Neste caso, o valor p é bastante superior ao nível de significância de 5% (0,05), logo a hipótese nula não pode ser rejeitada, indicando que os dados podem ser considerados normalmente distribuídos.

De forma a decidir que teste de hipótese aplicar verificamos as condições:

- n < 30,
- X $\cap$ Normal
- σ desconhecido

Aplicamos o teste de hipóteses – **t.test**:

Valor-p = 0.0058

Como o valor $p < \alpha$ então rejeita-se H0 num nível de significância de 5%, concluindo que a hipótese alternativa – H1: O valor médio dos custos iniciais para os indivíduos com idade inferior ou igual a 45 anos é significativamente superior a 3 (ou seja, H1: $\mu>3$) - é provavelmente verdadeira.

Conclusão: O valor médio dos custos iniciais para os indivíduos com idade inferior ou igual a 45 anos é significativamente superior a 3.

11. Verifique se existem diferenças significativas entre o valor médio do poder de compra (Poder_compra) entre homens e mulheres.

Para verificar se existem diferenças significativas entre o valor médio do poder de compra entre homens e mulheres, aplica-se um teste de hipóteses.

Conforme analisado anteriormente, subdividindo os homens e as mulheres verifica-se que em ambos os subconjuntos n > 30. Neste sentido, não é necessário verificar a normalidade da distribuição dos dados.

Condições de aplicabilidade:

Nível de significância – 5% ($\alpha = 0,05$)

Formulação das hipóteses:

- *Hipótese nula* (H0): Não há diferença significativa no valor médio do poder de compra entre homens e mulheres.
- *Hipótese alternativa* (H1): Há diferença significativa no valor médio do poder de compra entre homens e mulheres.

De forma a decidir que teste de hipóteses aplicar, verificamos as condições:

- n > 30
- σ desconhecido

Aplicamos o teste de hipóteses – **t.teste**:

Vamos realizar um t-teste para as duas amostras independentes para verificar se as médias são estatisticamente diferentes. O t-teste de duas amostras (two sample t-test) é um teste estatístico utilizado para comparar as médias de duas amostras independentes. Determina se há uma diferença estatisticamente significativa entre as médias das duas populações das quais as amostras foram retiradas.

Resultado t-teste: t=-1.295, valor-p=0.198 (>0.05)

Como o valor p > α então não se rejeita H0 para um nível de significância 5%, concluindo que os dados não fornecem evidencia para apoiar H1.

Concluindo-se que não se rejeita a hipótese nula de que não há diferença significativa no valor médio do poder de compra entre homens e mulheres, não havendo evidência nos dados que suportem a hipótese alternativa.

- **Conclusão:** Não há diferença significativa no valor médio do poder de compra entre homens e mulheres.

12. Utilizando a base de dados de que dispõe proceda a uma análise estatística suplementar elaborando 3 questões e tire conclusões.

12.1. Compare o custo médio da primeira receita entre tipos de hipertensão e represente-o graficamente.

Com recurso ao R Studio, calculamos o custo médio da primeira receita para cada um destes subconjuntos criados por tipo de hipertensão.

***Tabela 11**- Custo inicial médio por tipos de hipertensão*

	K86 – hipertensão sem complicações	K87 – hipertensão com complicações
$\bar{x}$ **Custo inicial (€)**	5.65	4.29

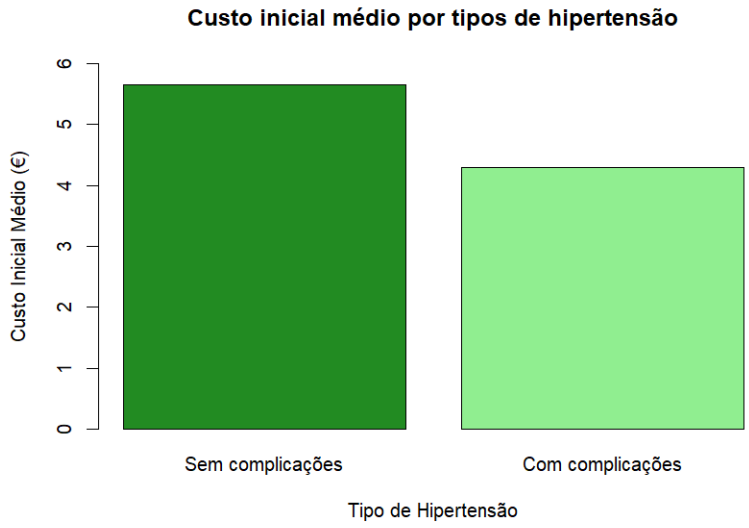


***Gráfico 26**- Gráfico de Barras - Custo inicial médio por tipos de hipertensão*

Através do gráfico de barras (gráfico 26), visualizamos que se verificam maiores encargos nos custos iniciais nos doentes com hipertensão sem complicações do que nos com complicações, ainda que sem grande expressão, contrariamente ao esperado observar se considerássemos também neste estudo as comorbilidades que possam estar associadas nos tipos de hipertensão em estudo.

12.2 Verifique se existem diferenças significativas entre o valor médio do custo da primeira receita e o custo da receita passada 24 meses depois.

Para verificar se existem diferenças significativas entre o valor médio do custo da primeira receita e o valor médio do custo da receita passada 24 meses depois, aplica-se um teste de hipóteses.

Para verificar as condições de aplicabilidade, confirmamos o tamanho da amostra para as variáveis em estudo, Custo_inicial e Custo_24. Conforme verificado em R através da função length, para ambas a amostra n é superior a 30 (n = 101).

Estamos assim a fazer uma comparação pareada. Um teste de hipóteses pareado é uma técnica estatística usada para comparar duas médias quando as observações nas duas amostras estão emparelhadas ou relacionadas de alguma forma. Por exemplo, aplica-se a estudos onde o mesmo indivíduo é medido duas vezes, com é o caso dos doentes deste estudo. Isto significa que cada observação na amostra do custo inicial corresponde a uma observação na amostra do custo após 24 meses.

O objetivo é testar a hipótese formulada (a hipótese de o valor médio do custo da primeira receita e o custo da receita passada 24 meses depois serem significativamente diferentes). A esta hipótese damos o nome de H1 (Hipótese Alternativa). Esta hipótese será testada contra a hipótese de acontecer o oposto, a que chamamos de H0 (Hipótese Nula).

O teste de hipóteses obtém-se a partir de um t-test, neste caso t.test pareado, no qual se obtêm um valor-p que é depois comparado com o nível de significância previamente fixado.

Condições de aplicabilidade teste hipóteses:

- Nível de significância – 5% ($\alpha = 0,05$)
- $n > 30$ (logo, não precisamos de verificar a normalidade dos dados)
- σ desconhecido

Formulação das hipóteses:

Hipótese nula (H0): O valor médio entre o custo médio da 1ª receita e o custo médio da receita 24 meses depois é significativamente idêntico. (H0: custo médio 1ª receita = custo médio receita 24 meses depois)

Hipótese alternativa (H1): O valor médio entre o custo médio da 1ª receita e o custo médio da receita 24 meses depois é significativamente diferente. (H1: custo médio 1ª receita ≠ custo médio receita 24 meses depois)

Aplicamos o teste de hipóteses - **t-teste pareado**:

Resultado t-teste: valor-p<.001

O valor-p <.001 indica que a probabilidade de observar os resultados do teste, ou resultados mais extremos, sob a hipótese nula é muito baixa. Portanto, com base no valor-p < 0.001, concluímos que há uma diferença estatisticamente significativa entre o custo médio da primeira receita e o custo médio da receita 24 meses depois.

12.3 Verifique se existe associação entre o sexo e o início do tratamento e tire conclusões.

Para dar resposta a esta questão é necessário caracterizar as variáveis a analisar para aplicar os métodos estatísticos adequados.

Variáveis a avaliar – Sexo, início – variáveis qualitativas nominais, ambas dicotómicas (ou seja, apresentam apenas duas classes em cada categoria).

Os métodos estatísticos adequados serão:

- Tabela de contingência
- Coeficiente de associação: Coeficiente Phi

Para analisarmos a associação entre estas duas variáveis construindo-se uma tabela de contingência de frequência relativas (em percentagem) através da função *prop.table*.

A tabela de contingência auxilia a organização e visualização dos dados em relação as variáveis qualitativas, como é o caso, permitindo uma melhor interpretação. Este passo é necessário para o cálculo do coeficiente de associação, a partir da tabela de contingência.

***Tabela 12**- Tabela de frequências relativas (sexo, início)*

	Não inicia tratamento (%)	Inicia tratamento (%)	Total (%)
Masculino (%)	6.93	32.67	39.6
Feminino (%)	4.95	55.45	60.4
Total (%)	11.88	88.12	100

Os coeficientes de associação são medidas estatísticas que refletem a relação entre duas variáveis qualitativas nominais. O coeficiente Phi é usado para avaliar a força da relação entre as variáveis em estudo (variáveis qualitativas dicotómicas).

O coeficiente de associação Phi apresenta uma distinta vantagem em relação aos outros coeficientes pelo facto de, para além de avaliar a força da ligação, avalia também o sentido da associação, ou seja, se as variáveis possuem uma associação no mesmo sentido ou no sentido oposto. Este coeficiente apresenta um domínio entre -1 e 1(inclusivamente) pelo que este valor

quando muito próximo de 0 demonstra que não existe associação; valores negativos próximos de -1 indicam associação inversa; e valores positivos próximos de 1 indicam associação direta.

Através da função *assocstats*, obtém-se o valor do coeficiente Phi = 0.141, o que reflete pouca ou nenhuma associação entre o sexo e o início do tratamento.

Referências Bibliográficas

Afonso, A., & Nunes, C. (2019). *Probabilidades e Estatística - Aplicações e Soluções em SPSS*. Universidade de Évora.

Hall, A., Neves C., & Pereira A. (2006). Grande Maratona de Estatística no SPSS. Capítulo 4.2. Testes de Hipóteses Paramétricos. Disponível em http://www2.mat.ua.pt/pessoais/AHall/me/files/TH2006.pdf

Mailund, T. (2017). *Beginning Data Science in R - Data Analysis, Visualization, and Modelling for the Data Scientist*. Apress.