## 4.6 Análise de Variância (ANOVA)

Na Seção 4.5 foi visto como realizar testes de hipóteses quando o interesse é comparar duas populações com relação às suas médias. No entanto, se o interesse for comparar mais de duas populações com relação às suas médias uma Análise de Variância em uma via (ANOVA-oneway) pode ser realizada. Ou seja, o interesse é verificar se uma variável qualitativa com mais de dois níveis tem efeito sobre uma variável quantitativa. Portanto, assumindo $k$ populações, o objetivo será testar as seguintes hipóteses:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k \\ H_1 : \text{pelo menos uma das médias é diferente das demais} \end{cases} \tag{4.20}$$

Para ilustrar o uso da ANOVA-oneway vamos considerar o conjunto de dados "anova.xls" (disponível na plataforma Moodle). Esse conjunto de dados se trata do tempo de durabilidade, em semanas, medidos em três tipos (marcas) de rolamentos utilizados em máquinas de uma linha de produção. Foram selecionados 14 rolamentos de cada tipo (3 marcas) e medido o tempo de duração dos rolamentos em semanas, totalizando $3 \times 14 = 42$ observações. Supor o interesse em verificar o efeito de tipo de rolamento (Tipo 1, Tipo 2 e Tipo Padrão) no tempo de duração. Seguindo os passos vistos na Seção 4.3 para a construção de testes de hipóteses, tem-se:

1. **Identificar $H_0$ e $H_1$**: Como o interesse é verificar se o tempo de duração é afetado pelo tipo de rolamento, tem-se o interesse em testar as seguintes hipóteses,

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_B \\ H_1 : \text{pelo menos uma das médias é diferente das demais} \end{cases}, \tag{4.21}$$

   em que, $\mu_1$ é a média do tempo de duração dos rolamentos do Tipo 1; $\mu_2$ é a média do tempo de duração dos rolamentos do Tipo 2 e $\mu_B$ é a média do tempo de duração dos rolamentos do Tipo Padrão (Baseline).

2. **Escolher o teste estatístico**: Como o interesse aqui é testar a igualdade entre mais de duas médias (3 médias), o teste que será realizado é a **ANOVA-oneway**.

3. **Fixar o nível de significância** $\alpha$: Cometer o erro do Tipo I aqui é dizer que, de acordo com a amostra pelo menos uma das médias do tempo de duração para os três tipos de rolamentos é diferente das demais, quando na população as médias do tempo de duração dos rolamentos é a mesma para os 3 tipos de rolamentos. Supor que o grau de gravidade em cometer esse erro é brando, o nível de significância aqui será estipulado em $\alpha = 10\%$.

   Os pressupostos desse teste são três, os resíduos ($\varepsilon$) gerados pelo modelo da ANOVA-oneway devem seguir distribuição normal com média zero, devem ser homocedásticos (variância constante) e devem ser independentes, ou seja:

$$\varepsilon \overset{i.i.d}{\sim} N\left(0, \sigma^2\right). \tag{4.22}$$

   Nesse texto não iremos abordar a fundo o que são esses resíduos gerados pela ANOVA-oneway. O importante é ter em mente que a ANOVA-oneway geram esses resíduos, que são facilmente obtidos utilizando o `Software R`, e que para a ANOVA-oneway ter validade os pressupostos acima devem ser observados.

   Para gerar os resíduos da ANOVA-oneway utilizando o `Software R`, é necessário, primeiramente, realizar os cálculos matemáticos envolvendo a ANOVA-oneway utilizando o comando `aov`, como segue (aqui foi criado o objeto "dados" contendo a matriz de dados):

ANOVA <- aov(dados$Durabilidade∼dados$Tipo)

O primeiro argumento da função `aov` é separado pelo símbolo ∼, antes do símbolo ∼ é definido o vetor com as observações da variável quantitativa em que se tem o interesse em realizar a ANOVA-oneway, após o símbolo ∼ é definido o vetor com as observações da variável qualitativa que representa os grupos que serão comparados. Veja que aqui foi criado o objeto "ANOVA" que contém todos os resultados matemáticos envolvendo a ANOVA-oneway. A partir do objeto "ANOVA" basta utilizar a sintaxe `ANOVA$res` para obter o vetor com os resíduos do modelo da ANOVA-oneway. A partir desse vetor é possível verificar se os resíduos seguem distribuição normal.

Portanto, de acordo com o gráfico da Figura 4.7 percebe-se que eles estão seguindo um comportamento linear acompanhando a linha vermelha teórica e como o p-valor do teste de Shapiro-Wilk, $p - valor = 0,8625$, é maior do que o nível de significância do teste ($\alpha = 10\%$), a hipótese de normalidade dos resíduos não é rejeitada.

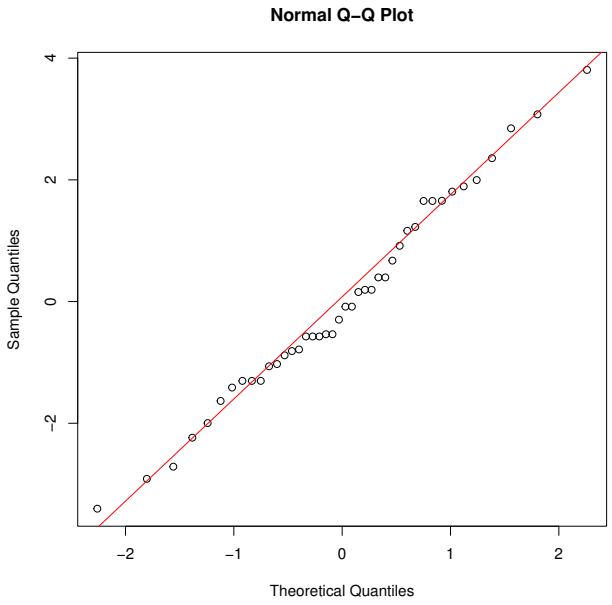



Figura 4.7: Gráficos quantil-quantil para verificar se os resíduos da ANOVA-oneway segue distribuição normal.

Para verificar a homoscedasticidade dos resíduos dois testes de hipóteses podem ser utilizados, o teste de homoscedasticidade de Breusch-Pagan e o teste de homoscedasticidade de Goldfeld-Quandt. As hipóteses testadas por esses testes são:

$$\begin{cases} H_0 : \text{os resíduos do modelo são homocedásticos} \\ H_1 : \text{os resíduos do modelo são heterocedásticos} \end{cases}. \tag{4.23}$$

Para realizar esses testes de hipóteses utilizando o `Software R`, primeiramente o pacote `lmtest` deve ser instalado (`install.packages("lmtest")`) e carregado (`library(lmtest)`) no software, depois basta executar as seguintes linhas de comando:

```
bptest(ANOVA)
gqtest(ANOVA)
```

O argumento das funções `bptest` e `gqtest` é o comando contendo os resultados da ANOVA-oneway gerados a partir da função `aov` (argumento "ANOVA" criado acima). O p-valor do teste de Breusch-Pagan é de $p - valor = 0,1072$ e o p-valor do teste de Goldfeld-Quandt é de $p - valor = 0.5262$. Como esses p-valores são maiores do que o nível de significância do teste ($\alpha = 10\%$), a hipótese de homoscedasticidade dos resíduos não é rejeitada.

Para verificar a independência dos resíduos o teste de independência de Durbin-Watson pode ser utilizado. As hipóteses testadas por esse teste são:

$$\begin{cases} H_0 : \text{os resíduos do modelo são independentes} \\ H_1 : \text{os resíduos do modelo são dependentes} \end{cases} . \tag{4.24}$$

Para realizar esses testes de hipóteses utilizando o `Software R`, primeiramente o pacote `lmtest` deve ser instalado (`install.packages("lmtest")`) e carregado (`library(lmtest)`) no software, depois basta executar a seguinte linha de comando:

```
dwtest(ANOVA)
```

O argumento da função `dwtest` é o comando contendo os resultados da ANOVA-oneway gerados a partir da função `aov` (objeto "ANOVA" criado acima). O p-valor do teste de Durbin-Watson é de $p - valor = 0,7817$, como esse p-valor é maior do que o nível de significância do teste ($\alpha = 10\%$), a hipótese de independência dos resíduos não é rejeitada.

Como todos os pressupostos da ANOVA-oneway forma observados ($\varepsilon \overset{i.i.d}{\sim} N(0, \sigma^2)$), então esse teste pode ser utilizado para resolver o problema.

4. **Calcular os valores observados para o teste estatístico a partir dos dados amostrais**: Para obter o p-valor da ANOVA-oneway é necessário utilizar os resultados obtidos pelo comando `aov` (objeto "ANOVA" criado no Item 3) dentro da função `summary` do `Software R`. Para o software retornar apenas o p-valor que testam as hipóteses dadas em (4.21), a seguinte linha de comando deve ser executada:

```
summary(ANOVA)[[1]][["Pr(>F)"]]
```

Executando essa linha de comando o `Software R` irá retornar o valor 0.001464939.

5. **Verificar se rejeita ou não a hipótese nula** $H_0$: Observando o resultado acima, o p-valor é dado por $p - valor = 0,00146$, como o nível de significância estipulado foi de $\alpha = 0,10$, então $p - valor < \alpha$. Portanto, com 10% de significância, existem evidências de que pelo menos uma das médias do tempo de duração dos rolamentos é diferente das demais, ou seja, existe um efeito do tipo de rolamento no tempo de duração.

   Como visto acima, pelo menos uma das médias do tempo de duração dos rolamentos é diferente das demais, no entanto não se sabe quais médias se diferem entre si. Para verificar isso é necessário realizar o chamado pós-teste, o pós-teste utilizado aqui será o pós-teste de Tukey HSD (Honest Significant Difference). Para realizar esse teste no `Software R` será utilizado a função `TukeyHSD`, como segue:

```
TukeyHSD(ANOVA,conf.level=0.90)
```

O primeiro argumento da função `TukeyHSD` é o comando contendo os resultados da ANOVA-oneway gerados a partir da função `aov` (objeto "ANOVA" criado no Item 3); o argumento `conf.level` define o nível de confiabilidade estipulado para o cálculo do intervalo de confiança. Executando essa linha de comando o `Software R` irá retornar o seguinte resultado:

Tukey multiple comparisons of means

90% family-wise confidence level

Fit: aov(formula = dados$Durabilidade ~ dados$Tipo)

$`dados$Tipo`

| | diff | lwr | upr | p adj |
|---|---|---|---|---|
| Type 1-Baseline | 2.530000 | 1.16781191 | 3.8921881 | 0.0009730 |
| Type 2-Baseline | 1.423571 | 0.06138333 | 2.7857595 | 0.0820676 |
| Type 2-Type 1 | -1.106429 | -2.46861667 | 0.2557595 | 0.2116114 |

Observando a primeira coluna dos resultados acima conjuntamente com a coluna `p adj`, temos as seguintes conclusões: $\mu_1 \neq \mu_B$, pois o p-valor observado, $p - valor = 0,0009730$, é menor do que o nível de significância estipulado $\alpha = 0,10$, ou seja, com 10% de significância pode-se dizer que o tempo médio de duração dos rolamentos do Tipo 1 é diferente do tempo médio de duração dos rolamentos do Tipo Padrão; $\mu_2 \neq \mu_B$, pois o p-valor observado, $p - valor = 0,0820676$, é menor do que o nível de significância estipulado $\alpha = 0,10$, ou seja, com 10% de significância pode-se dizer que o tempo médio de duração dos rolamentos do Tipo 2 é diferente do tempo médio de duração dos rolamentos do Tipo Padrão; $\mu_2 = \mu_1$, pois o p-valor observado, $p - valor = 0,2116114$, é maior do que o nível de significância estipulado $\alpha = 0,10$, ou seja, com 10% de significância pode-se dizer que o tempo médio de duração dos rolamentos do Tipo 2 é igual ao tempo médio de duração dos rolamentos do Tipo 1.

Para quantificar as diferenças observamos as colunas `lwr` e `upr` dos resultados acima. Esses valores são, respectivamente, os limites inferiores e superiores dos intervalos de confiança para a diferença entre as médias. Logo, temos que $IC\left(\mu_1 - \mu_B, 90\%\right) = [1,17; 3,89]$, ou seja, com 90% de confiabilidade, levantamos evidências de que o tempo médio de duração dos rolamentos do Tipo 1 é de $1,17$ semanas a $3,89$ semanas maior, se comparado com os rolamentos do Tipo Padrão. Temos também que $IC\left(\mu_2 - \mu_B, 90\%\right) = [0,06; 2,79]$, ou seja, com 90% de confiabilidade, levantamos evidências de que o tempo médio de duração dos rolamentos do Tipo 2 é de $0,06$ semana a $2,79$ semanas maior, se comparado com os rolamentos do Tipo Padrão.

Outro fato importante que podemos observar está relacionado ao resultado: $IC\left(\mu_2 - \mu_1, 90\%\right) = [-2.47; 0.26]$. Os limites desse intervalo de confiança contém o valor zero, isso implica que $\mu_2 - \mu_1 = 0 \Rightarrow \mu_1 = \mu_2$ (hipótese nula não rejeitada).

## 4.7 Teste de Correlação de Pearson

Frequentemente, tem-se o interesse em determinar se existe alguma correlação entre duas variáveis aleatórias quantitativas $X$ e $Y$, o que pode ser feito pelo teste de hipóteses,

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}; \tag{4.25}$$

ou seja, se a hipótese nula não for rejeitada, então, com um certo nível de significância, pode-se dizer que as duas variáveis aleatórias quantitativas não estão correlacionadas. No entanto, se a hipótese nula for rejeitada, então, com um certo nível de significância, pode-se dizer que as duas variáveis aleatórias quantitativas estão correlacionadas. O teste de hipótese utilizado para testar essas hipóteses é o **Teste de Correlação de Pearson**. O pressuposto desse teste é que as duas variáveis aleatórias que serão correlacionadas devem seguir distribuição normal.

Para ilustrar o uso do teste de correlação de Pearson vamos considerar o conjunto de dados "Dieta.xlsx" (disponível na plataforma Moodle). Supor o interesse em verificar se a proporção de perda de peso esta correlacionada com a idade dos indivíduos submetidos as duas dietas. Seguindo os passos vistos na Seção 4.3 para a construção de testes de hipóteses, tem-se:

1. **Identificar** $H_0$ **e** $H_1$: Como o interesse é verificar se a proporção de perda de peso esta correlacionada com a idade dos indivíduos submetidos as duas dietas, tem-se o interesse em testar as seguintes hipóteses,

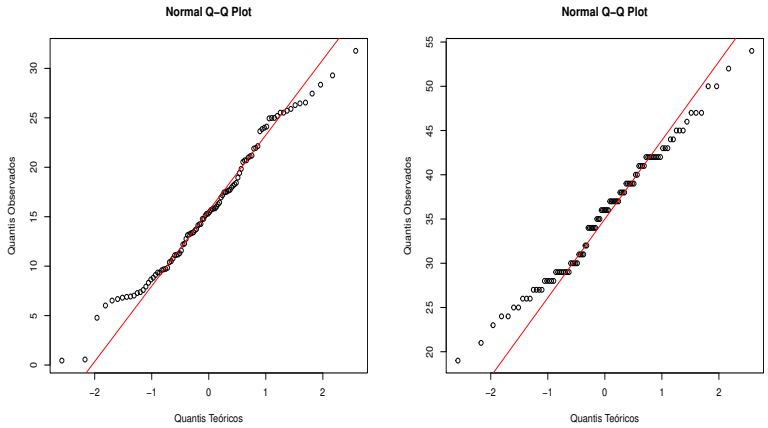
$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}, \tag{4.26}$$

   ou seja, tem-se o interesse em verificar se o coeficiente de correlação $\rho$ é igual ou diferente de zero. Se o coeficiente de correlação for diferente de zero há indícios de que as duas variáveis quantitativas estão correlacionadas, caso contrário ($\rho = 0$) as duas variáveis quantitativas não estão correlacionadas.

2. **Escolher o teste estatístico**: Como o interesse aqui é testar a correlação entre duas variáveis quantitativas, o teste que será realizado é o **Teste de Correlação de Pearson**.

3. **Fixar o nível de significância** $\alpha$: Cometer o erro do Tipo I aqui é dizer que, de acordo com a amostra a proporção de perda de peso esta correlacionada com a idade dos indivíduos, quando na população as duas variáveis não estão correlacionada. Supor que o grau de gravidade em cometer esse erro é brando, o nível de significância aqui será estipulado em $\alpha = 10\%$.

   Lembrando que esse teste tem como pressuposto que as duas variáveis aleatórias que serão correlacionadas devem seguir distribuição normal. Portanto, a partir dos gráficos da Figura 4.8 percebe-se que eles estão seguindo um comportamento linear acompanhando a linha vermelha teórica e como os p-valores do teste de Shapiro-Wilk para as variáveis proporção de perda de peso e idade ($p - valor_P = 0,1717$ e $p - valor_I = 0,3137$), é maior do que o nível de significância do teste ($\alpha = 10\%$), a hipótese de normalidade dos dados não é rejeitada. Portanto, o teste de correlação de Pearson pode ser utilizado para resolver o problema.




(a) Proporção de perda de peso (b) Idade

Figura 4.8: Gráficos quantil-quantil para verificar se a proporção de perda de peso e a idade seguem distribuição normal para indivíduos submetidos a Dieta A e B.

4. **Calcular os valores observados para o teste estatístico a partir dos dados amostrais**: Para realizar o **Teste de Correlação de Pearson** utilizando o `Software R`, a função `cor.test` será utilizada e a seguinte linha de comando deve ser executada:

   cor.test(dados$Perda,dados$Idade,method="pearson",conf.level=0.90)

   Os dois primeiros argumentos da função `cor.test` são os vetores contendo os dados das variáveis aleatórias quantitativas em que se tem o interesse em verificar se há a correlação; no argumento `method` é definido o tipo de teste de correlação que será utilizado, nesse caso o teste é o de Pearson (outros testes de correlação serão abordados na próxima seção desse texto); o argumento `conf.level` define o nível de confiabilidade estipulado para o cálculo do intervalo de confiança. Executando essa linha de comando o `Software R` irá retornar o seguinte resultado:

   Pearson's product-moment correlation

   data: dados$Perda and dados$Idade

   t = -5.4784, df = 98, p-value = 3.323e-07

   alternative hypothesis: true correlation is not equal to 0

   90 percent confidence interval:

   -0.6014859 -0.3464932

   sample estimates:

   cor

   -0.484205

5. **Verificar se rejeita ou não a hipótese nula** $H_0$: Observando os valores acima, o p-valor é dado por $p-valor = 3.323e-07$, como o nível de significância estipulado foi de $\alpha = 0,10$, então $p - valor < \alpha$. Portanto, com 10% de significância, existem evidências de que a proporção de perda de peso esta correlacionada com a idade dos indivíduos.

O `Software R` também retorna o intervalo de confiança para o coeficiente de correlação de Pearson, logo, temos que $IC(\rho, 90\%) = [-0,60; -0,35]$. Portanto, com 90% de confiabilidade, levantamos evidências de que há uma correlação negativa e fraca (ver Tabela 2.3), em que, quanto maior a idade do indivíduo menor a sua proporção de perda de peso.