

### 2.4.3 Gráficos para verificar normalidade

Um primeiro gráfico chamado de qq-norm permite a comparação de duas distribuições de probabilidades traçando seus quantis uns contra os outros. Depois exploramos um gráfico mais recente, conhecido como worm plot (gráfico de minhoca), consistindo numa determinada coleção de de qq-norm.

#### QQ-norm

O gráfico quantil-quantil ou qq-plot, proposto por Wilk & Gnanadesikan (1968), é um dispositivo gráfico exploratório utilizado para verificar a validade de um pressuposto de distribuição para um conjunto de dados. Em geral, a ideia básica é a de calcular o valor teoricamente esperado para cada ponto de dados com base na distribuição em questão. Se os dados de fato seguirem a distribuição assumida os pontos deste gráfico formarão aproximadamente uma linha reta.

Percebemos que podemos verificar com este gráfico qualquer densidade contínua, eventualmente pode ser utilizado também para funções de probabilidade. O qq-plot vai apresentar-se como uma linha reta se a densidade assumida estiver correta. Vejamos o caso particular de verificarmos se a densidade é normal, nesta situação o gráfico qq-plot será chamado de qq-norm. Primeiro consideraremos a situação da densidade normal padrão.

Seja  $z_1, z_2, \dots, z_n$  uma amostra aleatória de uma distribuição normal com média  $\mu = 0$  e desvio padrão  $\sigma = 1$ . As estatísticas de ordem amostrais são

$$z_{(1)} \leq z_{(2)} \leq \dots \leq z_{(n)}.$$

Estes valores desempenharão o papel dos quantis da amostra. Agora, quais devemos tomar como os quantis teóricas correspondentes? Se a função de distribuição cumulada da densidade normal padrão fosse denotada por  $\Phi$ , usando a notação quantil, se  $\xi_q$  é o  $q$ -ésimo quantil de uma distribuição normal, então

$$\Phi(\xi_q) = q,$$

ou seja, a probabilidade de uma amostra normal ser inferior a  $\xi_q$  é, de fato, apenas  $q$ .

Considere o primeiro valor ordenado  $z_{(1)}$ . O que podemos esperar que o valor  $\Phi(z_{(1)})$  seja? Intuitivamente, esperamos que essa seja a probabilidade de assumir um valor no intervalo  $(0, 1/n)$ . Do mesmo modo, espera-se que  $\Phi(z_{(2)})$  seja a probabilidade de assumir um valor no intervalo  $(1/n, 2/n)$ .

Continuando, esperamos que  $\Phi(z_{(n)})$  seja a probabilidade de assumir um valor no intervalo  $(n-1)/n, 1)$ . Assim, o quantil teórico desejamos seja definido pelo inverso da função de distribuição acumulada normal padrão. Em particular, o quantil teórico correspondente ao quantil empírico  $z_{(i)}$  deve ser

$$\xi_q = q \approx \frac{i - 0,5}{n},$$

para  $i = 1, 2, \dots, n$ .

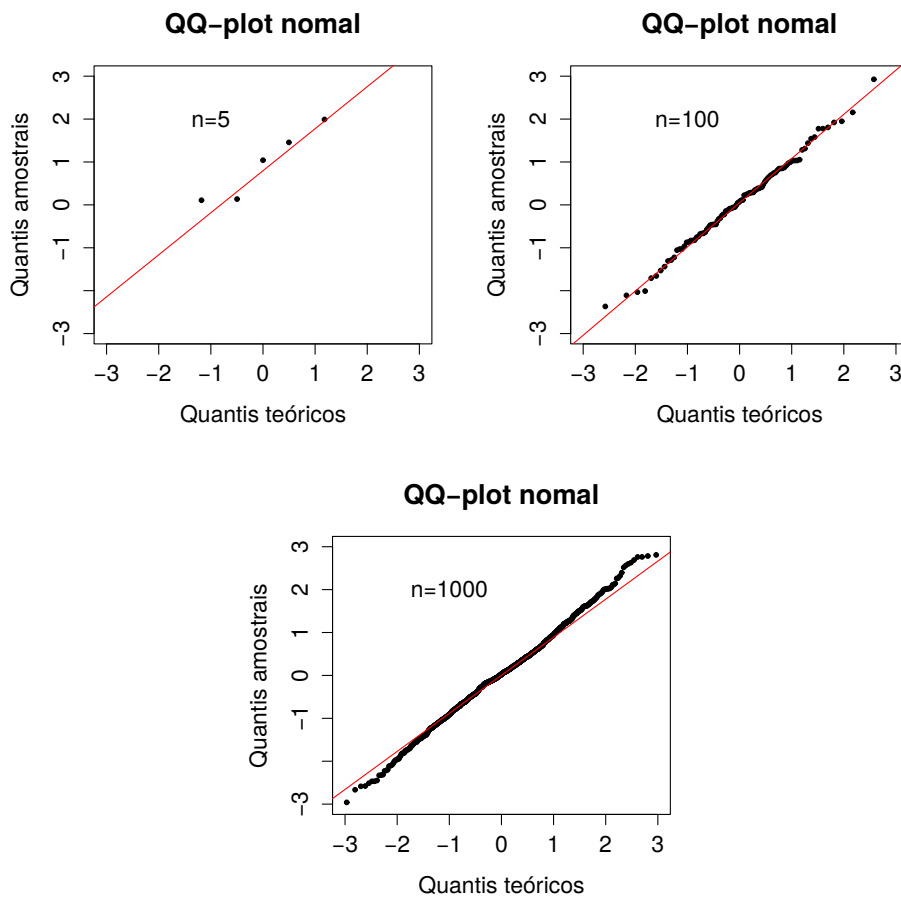


Figura 2.11: Diferentes qqplot para dados normais.

Na Figura 2.11, a esquerda acima exibimos o qq-norm de uma pequena amostra normal de tamanho 5. Os restantes quadros na Figura 2.11 exibem

as plotagens de qq-norm para amostras normais de tamanhos  $n = 100$  e  $n = 1000$ , respectivamente. Como o tamanho da amostra aumenta, os pontos encontram-se mais perto da linha  $y = x$ .

Estes gráficos (Figura 2.11) foram gerados utilizando as linhas de comando:

```
set.seed(1278)
x=rnorm(5)
qqnorm(x, xlim=c(-3,3), ylim=c(-3,3), cex=0.6, pch=19,
        ylab='Quantis amostrais', xlab='Quantis teóricos',
        main='QQ-plot normal')
qqline(x,col="red")
text(-1,2,'n=5')
```

para a situação de amostra de tamanho 5. A primeira linha de comando serve para fixar o gerador de números aleatórios e, dessa forma, podemos simular sempre a mesma amostra e reproduzir o gráfico idêntico. Nas outras situações somente muda-se o tamanho da amostra que se quer gerar. Assim, os comandos para gerar o segundo e terceiro gráficos são, respectivamente

```
x=rnorm(100)
qqnorm(x, xlim=c(-3,3), ylim=c(-3,3), cex=0.6, pch=19,
        ylab='Quantis amostrais', xlab='Quantis teóricos',
        main='QQ-plot normal')
qqline(x,col="red")
text(-1,2,'n=100')
```

e

```
x=rnorm(1000)
qqnorm(x, xlim=c(-3,3), ylim=c(-3,3), cex=0.6, pch=19,
        ylab='Quantis amostrais', xlab='Quantis teóricos',
        main='QQ-plot normal')
qqline(x,col="red")
text(-1,2,'n=1000')
```

Caso os dados não forem padronizados basta aplicar a transformação  $(X - \hat{\mu})/\hat{\sigma}$ , onde  $X$  representa os dados originais e  $\hat{\mu}$  e  $\hat{\sigma}$  representam os estimadores dos parâmetros  $\mu$  e  $\sigma$ , respectivamente.

Estes gráficos podem indicar afastamentos da normalidade por isso apresentamos duas situações de dados não simétricos e com caudas pesadas. Na Figura 2.12, mostramos o que acontece se os dados forem da distribuição t-Student(8) e da distribuição  $\chi^2(5)$ , sempre de tamanho  $n = 1000$ . Observe, em particular, que os dados a partir da distribuição t-Student seguem

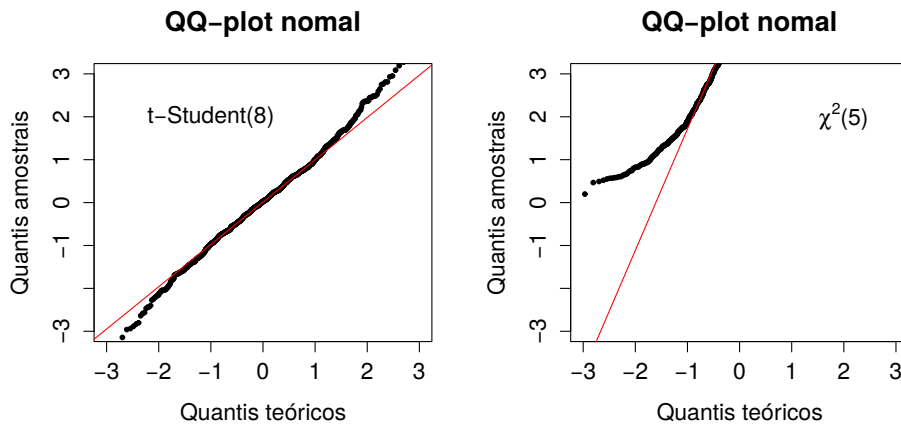


Figura 2.12: Diferentes qqplot para dados não normais.

a curva normal bem de perto até os últimos pontos em cada extremo. Na outra situação o afastamento da distribuição normal é evidente.

Foi mencionado que o qq-norm é uma situação particular do qq-plot devido a este último permitir comparar os quantis amostrais com os quantis distribucionais. Com isto queremos dizer que o qq-plot serve para verificar se os dados forem t-Student ou  $\chi^2(5)$ , por exemplo. Na Figura 2.13 apresentamos a aparência dos gráficos qq-plot caso queira-se verificar se as amostras seguem distribuição t-Student(8) ou  $\chi^2(5)$ , respectivamente.

Os gráficos na Figura 2.13 foram gerados pelas linhas de comandos

```
qqplot(qt(ppoints(1000), df = 8), x, cex=0.6, pch=19,
       main = "QQ plot para t-Student(8)",
       xlab="t-Student(8)")
qqline(x, distribution = function(p) qt(p, df = 8),
       prob = c(0.1, 0.6), col = 2)
```

no caso t-Student(8) e

```
qqplot(qchisq(ppoints(1000), df = 5), x, cex=0.6, pch=19,
       main = expression("QQ plot para" ~ {chi^2}(5)),
       xlab=expression({chi^2}(5)))
qqline(x, distribution = function(p) qchisq(p, df = 5),
       prob = c(0.1, 0.6), col = 2)
```

para o caso  $\chi^2(5)$ .

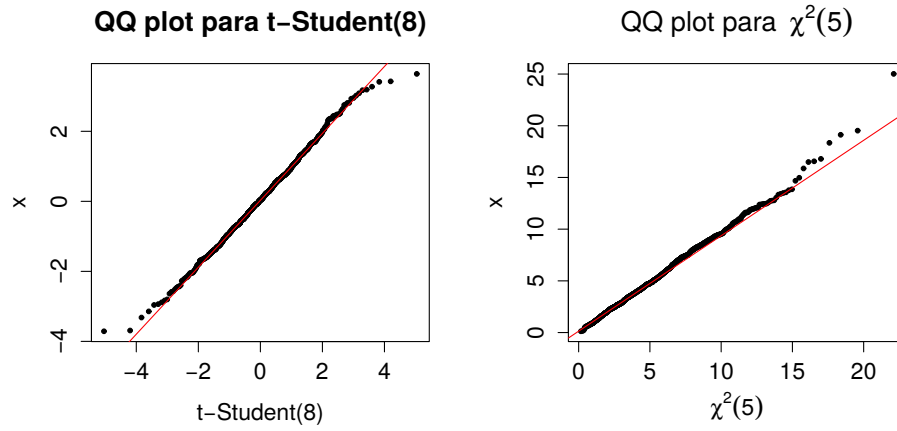


Figura 2.13: Diferentes qqplot para dados não normais.

### Worn plot

O worm-plot é uma série de parcelas de gráficos qq-plot retificados. Constitui uma ferramenta de diagnóstico para visualização de quão bem um modelo estatístico se ajusta aos dados, para encontrar locais em que o ajuste pode ser melhorado e para comparar o ajuste de diferentes modelos.

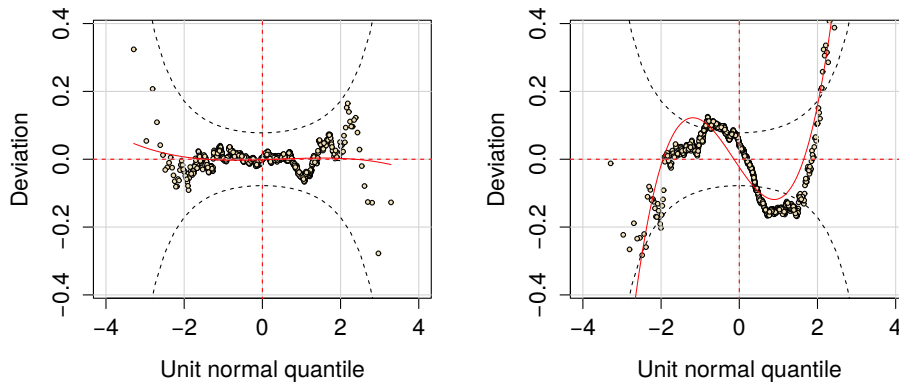


Figura 2.14: Diferentes worm-plot para dados normais.

Na Figura 2.14 mostramos este gráfico para duas situações: a esquerda

os dados são normais e a direita os dados são t-Student com 8 graus de liberdade. Nesta situação aparece bem a qualidade da observação com esta figura. Se os dados forem normais o curva worm-plot ou gráfico de minhoca deve aparentar um verme achatado, os pontos próximos a curva vermelha e com poucas oscilações. Quando aplicamos este gráfico ao caso t-Student percebemos uma oscilação grande no verme e com pontos fugindo da banda de confiança. Isso comprova que os dados não seguem como referência a distribuição normal. Agora, na Figura 2.15, a esquerda temos o caso de dados com distribuição  $\chi^2(5)$  e a direita dados com distribuição Cauchy padrão. Nestas situações fica claro que os dados não são normais.

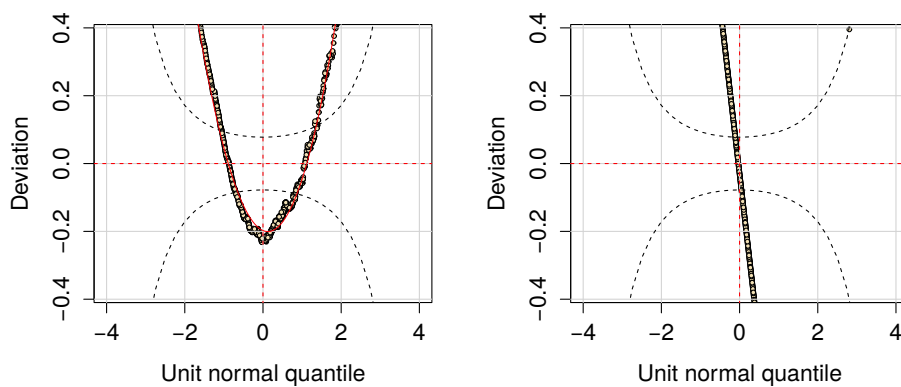


Figura 2.15: Diferentes worm-plot para dados não normais.

As linhas a seguir mostram os comandos necessários para gerar os gráficos nas Figuras 2.14 e 2.15. Utilizamos a biblioteca de comandos R `gamlss` (Rigby & Stasinopoulos, 2005).

```
library(gamlss)
x=rnorm(1000)
wp(gamlss(x~1), cex=0.6)
x=rt(1000, df=8)
wp(gamlss(x~1), cex=0.6)
x=rchisq(1000, df=5)
wp(gamlss(x~1), cex=0.6)
x=rcauchy(1000)
wp(gamlss(x~1), cex=0.6)
```