

# Estatística não paramétrica

## Testes de bondade de ajuste

Fernando Lucambio

Departamento de Estatística  
Universidade Federal do Paraná

20 de março de 2024

Um problema importante na estatística diz respeito à obtenção de informações sobre a forma da população da qual uma amostra é retirada. A forma dessa distribuição pode ser o foco da investigação. Alternativamente, algumas inferências sobre um aspecto particular da população podem ser de interesse. Neste último caso, na estatística clássica, as informações sobre a forma geralmente devem ser postuladas ou incorporadas na hipótese nula para realizar um tipo de inferência paramétrico exato.

Aqui vamos considerar dois tipos de testes de bondade de ajuste. O primeiro tipo é projetado para hipóteses nulas relativas a uma distribuição discreta e compara as frequências observadas com as frequências esperadas sob a hipótese nula. O segundo tipo de teste de bondade de ajuste é projetado para hipóteses nulas relativas a uma distribuição contínua e compara as frequências relativas acumuladas observadas com aquelas esperadas sob as hipóteses nulas.

Uma única amostra aleatória de tamanho  $n$  é extraída de uma população com função de distribuição desconhecida  $F$ . Desejamos testar a hipótese nula

$$H_0 : F(x) = F_0(x), \quad \text{para todo } x,$$

onde  $F_0(x)$  é completamente especificada, contra a alternativa geral

$$H_0 : F(x) \neq F_0(x), \quad \text{para algum } x.$$

Quando a distribuição da população é completamente especificada pela hipótese nula, pode-se calcular as probabilidades de que uma observação aleatória seja classificada em cada uma das categorias escolhidas ou fixas.

Essas probabilidades multiplicadas por  $n$  fornecem as frequências para cada categoria que seriam esperadas se a hipótese nula fosse verdadeira. Excetuando-se a variação amostral, deve haver uma concordância próxima entre essas frequências esperadas e observadas se os dados da amostra forem compatíveis com a distribuição  $F_0$  especificada.

As frequências observadas e esperadas correspondentes podem ser comparadas visualmente usando um histograma, um polígono de frequência ou um gráfico de barras. O teste qui-quadrado de bondade de ajuste fornece uma base probabilística para efetuar a comparação e decidir se a falta de concordância é muito grande para ter acurado por acaso.

A situação mais típica é quando a hipótese nula é composta, isto é, declaramos a forma da distribuição, mas não todos os parâmetros. Por exemplo, quando testamos se uma amostra é retirada de alguma população normal,  $\mu$  e  $\sigma$  não seriam dados.

No entanto, para calcular as frequências esperadas em  $H_0$ ,  $\mu$  e  $\sigma$  devem ser conhecidas. Se as frequências esperadas são estimadas a partir dos dados como  $n\hat{\theta}_i^0$ , a variável aleatória para o teste de ajuste assume a forma

$$Q = \sum_{i=1}^k \frac{(F_i - n\hat{\theta}_i^0)^2}{n\hat{\theta}_i^0}.$$

No teste de bondade de ajuste normal, por exemplo, as estimativas dos parâmetros  $\mu$  e  $\sigma$  seriam calculadas a partir dos dados agrupados e usadas com quantis da distribuição normal para encontrar  $n\hat{\theta}_i^0$  e os graus de liberdade para  $k$  categorias iria para  $k - 3$ .

Quando os dados originais são desagrupados e os estimadores de máxima verossimilhança são baseados na função de verossimilhança de todas as observações, a teoria é diferente. Chernoff and Lehmann (1954) mostraram que a distribuição limite de  $Q$  não é qui-quadrado neste caso e que  $P(Q > \chi_\alpha^2) > \alpha$ .

O teste é então não conservador. Na prática, a estatística é frequentemente tratada como uma variável qui-quadrado de qualquer maneira.

Neste teste procedemos da seguinte maneira: estimamos a função de distribuição empírica da amostra e calculamos uma estatística cujo valor mede o afastamento da função de distribuição empírica da amostra daquela da distribuição nula.

Seja  $X_1, X_2, \dots, X_n$  uma amostra aleatória da função de distribuição  $F$  e seja  $\hat{F}_n$  a correspondente função de distribuição empírica. A estatística

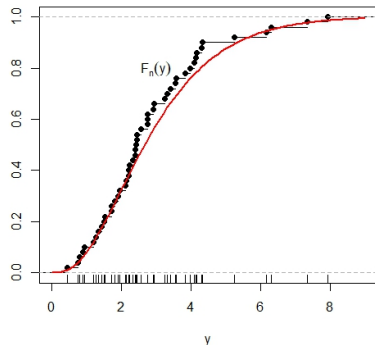
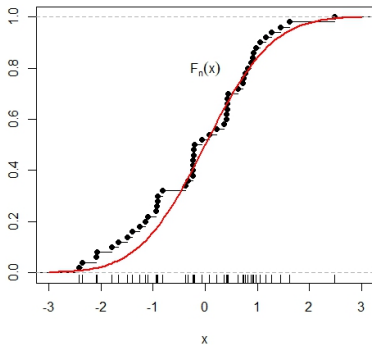
$$D_n = \sup_x |\hat{F}_n(x) - F(x)|,$$

é chamada de estatística de Kolmogorov-Smirnov bilateral. Escrevemos

$$D_n^+ = \sup_x (\hat{F}_n(x) - F(x)), \quad D_n^- = \sup_x (F(x) - \hat{F}_n(x)),$$

e chamamos  $D_n^+$  e  $D_n^-$  as estatísticas de Kolmogorov-Smirnov unilaterais.

Observemos na figura abaixo que o objetivo desta estatística é encontrar a diferença máxima entre a distribuição suposta dos dados  $F$  e a distribuição empírica  $\hat{F}_n$ . Nesta figura apresentamos dois exemplos: à esquerda simulamos 50 dados normais e mostramos a função de distribuição normal padrão acumulada, curva em vermelho, assim como a correspondente distribuição empírica; à direita simulamos uma distribuição diferente.





**Teorema:** As estatísticas  $D_n$ ,  $D_n^+$  e  $D_n^-$  são de distribuição livre para qualquer função de distribuição contínua  $F$ .

Gibbons (1971) encontrou a expressão de

$$P\left(D_n < \nu + \frac{1}{2n}\right),$$

se  $F$  for contínua.

Seja  $D_{n,\alpha}$  o ponto  $\alpha$ -percentual superior da distribuição de  $D_n$ , isto é,  $P(D_n > D_{n,\alpha}) \leq \alpha$ . A distribuição exata de  $D_n$  para valores selecionados de  $n$  e  $\alpha$  foi tabulada por Miller (1956), Owen (1962) e Birnbaum (1952). A distribuição  $D_n$  em amostras grandes foi derivada por Kolmogorov e a apresentamos a seguir.

**Teorema:** Seja  $F$  uma função de distribuição contínua. Então, para todo  $z \geq 0$

$$\lim_{n \rightarrow \infty} P(D_n \leq zn^{-1/2}) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 z^2}.$$

**Demonstração.** Ver Kolmogorov (1933).

Este teorema pode ser usado para encontrar  $d_n$  tal que

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n \leq d_n) = 1 - \alpha.$$

Tabelas de  $d_n$  para vários valores de  $\alpha$  foram disponibilizados por Owen (1962). As estatísticas  $D_n^+$  e  $D_n^-$  têm a mesma distribuição devido à simetria.

Tabelas para os valores críticos  $D_{n,\alpha}^+$ , onde

$$P(D_n^+ > D_{n,\alpha}^+) \leq \alpha$$

foram disponibilizados para valores selecionados de  $n$  e  $\alpha$  em Birnbaum and Tingey (1951).

No caso de amostras grandes Smirnov (1944) demonstrou que

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n^+ \leq z) = 1 - e^{-2z^2}, \quad z \geq 0.$$

Todos estes resultados estão programados na função R **ks.test**, esta função fornece-nos tanto o valor do estatística  $D_n$  quanto o p-valor, seja este encontrado de maneira exata ou através da distribuição limite.

O teste de normalidade de Lilliefors é uma adaptação do teste Kolmogorov-Smirnov para o caso em que a média e a variância da distribuição normal são desconhecidos. Seja  $X_1, X_2, \dots, X_n$  uma amostra aleatória de tamanho  $n$ . Estimamos a média e a variância dos dados usando os estimadores não viciados

$$\hat{\mu} = \bar{X} \quad \text{e} \quad \hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Aqui nós consideramos o problema do teste de bondade do ajuste para a distribuição normal sem média e variância especificadas. Este problema é muito importante na prática, porque a suposição de uma distribuição normal com  $\mu$  e  $\sigma$  é necessária para tantos testes estatísticos clássicos e procedimentos de estimação.

Nesse caso, observe que a hipótese nula é composta porque afirma que a distribuição subjacente é alguma distribuição normal. Em geral, o teste Kolmogorov-Smirnov pode ser aplicado no caso de hipóteses compostas de ajuste de qualidade após estimar os parâmetros desconhecidos,  $F_0$  será então substituído por  $\hat{F}_0$ .

infelizmente, a distribuição nula da estatística de teste Kolmogorov-Smirnov com parâmetros estimados é muito mais complicada. Isso, obviamente, afeta os cálculos do p-valor. Para a distribuição normal, Lilliefors (1967) mostrou que a utilização dos pontos críticos usuais desenvolvidos para o teste Kolmogorov-Smirnov fornece resultados extremamente conservadores. Ele então usou simulações Monte Carlo para desenvolver uma tabela para a estatística Kolmogorov-Smirnov que fornece valores críticos precisos.

Como antes, a estatística Kolmogorov-Smirnov é definida como

$$D_n = \sup_x |\hat{F}_n(x) - \hat{F}_0(x)|,$$

onde  $\hat{F}_n$  é a função de distribuição empírica e  $\hat{F}_0$  é calculado como a distribuição normal padrão  $\Phi(z)$  onde  $z = (x - \hat{x})/s$  para cada observação  $x$ ,  $\hat{x}$  a média amostral das  $n$  observações e  $s^2$  o estimador não viciado de  $\sigma^2$ , calculado com  $n - 1$  no denominador.

A hipótese nula de normalidade é rejeitada se  $D_n > c$ , onde  $c$  é o valor de corte. Os valores de corte desta estatística de teste para valores selecionados de  $n$  e diferentes níveis de significância podem ser encontrados em Dallal e Wilkinson (1986) sendo apenas confiáveis quando o p-valor é menor que 0.100.

Uma aproximação analítica, devida a estes mesmos autores, da probabilidade de cauda superior  $p$ , da distribuição de  $D_n$  para menos de 0.100 e tamanho de amostra  $n$  entre 5 e 100 é dada por

$$\exp \left( -7.01256D_n^2(n + 2.789019) + 2.99587D_n\sqrt{n + 2.78019} - 0.122119 + \frac{0.974598}{\sqrt{n}} + \frac{1.67997}{n} \right).$$

Para amostras de tamanho  $n > 100$ , a mesma expressão é utilizada com  $D_n$  substituído por  $D_n \times \left(\frac{n}{100}\right)^{0.49}$ . Se o p-valor de Dallal-Wilkinson for superior a 0.100, então o p-valor é calculado a partir da distribuição da estatística modificada  $D_n(\sqrt{n} - 0.01 + 0.85/\sqrt{n})$ , ver Stephens (1974).

A fórmula real do p-valor sendo obtida por um processo de simulação e aproximação disponível na função `lillie.test` no pacote **nortest**.

```
> dados = c(-1.787, -1.229, -0.525, -0.513, -0.508, -0.486, -0.482,  
            -0.323, -0.261, -0.068, -0.057, 0.137, 0.464, 0.595, 0.881,  
            0.906, 1.046, 1.237, 1.678, 2.455)  
> library(nortest)  
> lillie.test(dados)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: dados  
D = 0.149, p-value = 0.2899
```

Aceitamos a hipótese nula de normalidade dos dados.



A estatística do teste de Cramér-von Mises pertence ao grupo de estatísticas de bondade do ajuste baseadas em uma comparação da função de distribuição empírica de uma dada amostra com a distribuição teórica a ser testada. Ela é projetada para testar que a variável aleatória  $X$  tem uma distribuição contínua  $F(\cdot; \theta)$ ; sendo  $\theta$  um vetor de um ou mais parâmetros que entram na função de distribuição. Assim, para a distribuição normal, o vetor  $\theta = (\mu, \sigma^2)$ .

Muitas estatísticas deste tipo foram propostas, a mais famosa e uma das mais antigas sendo a  $D_n$  de Kolmogorov-Smirnov. Esta estatística é baseada na maior discrepância vertical entre as duas funções.

Uma medida alternativa é a família

$$W^* = \int_{-\infty}^{\infty} [\widehat{F}_n(x) - F(\cdot; \theta)]^2 \psi(x) f(x) dx,$$

sendo,  $f$  a função de densidade sob a hipóteses nula e, no caso da estatística Cramér-von Mises,  $\psi(x) = 1$ . A expressão computacional da estatística de teste é;

$$W_n^* = \frac{1}{n} \sum_{i=1}^n \left( F_0(X_{(i)}) - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n^2},$$

onde  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  são as estatísticas de ordem. Considerando certa a hipótese nula  $H_0$ , a distribuição limite desta estatística é bem conhecida e sabe-se que em amostras finitas  $W_n^*$  aborda este limite muito rapidamente.

**Exemplo.** Vamos considerar os seguintes dados para mostrar o comando que realiza o teste de bondade de ajuste de Cramer-von Mises para a distribuição especificada pelo argumento null. Assume-se que os valores em  $x$  sejam uma amostra aleatória com alguma função de distribuição  $F$ . A hipótese nula é que  $F$  seja a função especificada pelo argumento null, enquanto a hipótese alternativa é que  $F$  é alguma outra função.

```
> library(goftest)
> cvm.test(dados, null = "pnorm", mean = 0, sd = 1)
```

```
Cramer-von Mises test of goodness-of-fit
Null hypothesis: Normal distribution
with parameters mean = 0, sd = 1
```

```
data: dados
omega2 = 0.080897, p-value = 0.692
```

Esta estatística é projetada para testar que a variável aleatória  $X$  tem uma distribuição contínua  $F(\cdot; \theta)$ ; sendo  $\theta$  um vetor de um ou mais parâmetros que entram na função de distribuição. Assim, para a distribuição normal, o vetor  $\theta = (\mu, \sigma^2)$ .

Muitas estatísticas deste tipo foram propostas, a mais famosa e uma das mais antigas sendo a  $D_n$  de Kolmogorov-Smirnov. Esta estatística é baseada na maior discrepância vertical entre as duas funções. Uma medida alternativa é a família

$$W^* = \int_{-\infty}^{\infty} [\widehat{F}_n(x) - F(\cdot; \theta)]^2 \psi(x) f(x) dx,$$

sendo no caso desta situação

$$\psi(x) = \frac{1}{F(\cdot; \theta)[1 - F(\cdot; \theta)]}.$$

Essa função de peso neutraliza o fato de que a discrepância entre  $\hat{F}_n$  e  $F$  está necessariamente se tornando menor nas caudas, uma vez que ambas abordam 0 e 1 nos extremos. A função de peso dada pesa a discrepância por um fator inversamente proporcional à sua variância, e tem o efeito de dar grande importância às observações na cauda do que a maioria das outras estatísticas. Esta estatística é recomendada com boas propriedades de poder sobre uma ampla gama de distribuições alternativas quando  $F$  não é a verdadeira distribuição.

Para fins práticos, a definição da estatística Anderson-Darling acima precisa ser transformada em uma fórmula computacional.

Dada a amostra  $X_1, X_2, \dots, X_n$ , calculamos  $Z_i = F(X_{(i)}; \theta)$ ,  $i = 1, 2, \dots, n$  do qual temos que a estatística de Anderson-Darling assume a forma

$$A^2 = - \frac{\sum_{i=1}^n (2i - 1) \left( \ln(Z_i) - \ln(1 - z_{n+1-i}) \right)}{n} - n.$$

A fórmula para  $Z_i$  acima assume que a distribuição testada  $F$  é completamente especificada, isto é, os parâmetros em  $F$  devem ser conhecidos. A estatística  $A^2$  foi introduzida por Anderson and Darling (1954) e para esta situação eles deram a distribuição assintótica e tabelas de porcentagem de pontos. Para fins de teste, a cauda superior de  $A^2$  será usada; grandes discrepâncias entre a distribuição empírica e a distribuição testada indicarão um ajuste inadequado.

Mais tarde, Lewis (1961) demonstrou que a distribuição de  $A^2$  para uma amostra finita aproxima-se da distribuição assintótica de forma extremamente rápida de modo que, para fins práticos, apenas a distribuição assintótica é necessária para tamanho de amostra maior que 5.

A distribuição de  $A^2$  para o caso de parâmetros conhecidos é a mesma para todas as distribuições testadas. Isso ocorre porque a transformação integral da probabilidade é feita na primeira etapa, quando calculamos  $Z_i = F(X_{(i)}; \theta)$  e os valores de  $Z_i$  são valores ordenados de uma distribuição uniforme com os limites 0 e 1.  $A^2$  é, portanto, uma função de variáveis aleatórias uniformes ordenadas.

A teoria da distribuição assintótica para este caso especial pode ser encontrada a partir da teoria assintótica da função de distribuição empírica ou, mais especificamente, da função

$$\sqrt{n}(\hat{F}_n(z) - z).$$

Quando  $\theta$  contém componentes desconhecidos os  $Z_i$ , dados pela transformação acima, quando um estimador  $\hat{\theta}$  substitui  $\theta$ , não serão variáveis aleatórias ordenadas com distribuição uniformes e a teoria de distribuição de  $A^2$ , torna-se substancialmente mais difícil. Em geral, a distribuição de  $A^2$  depende de  $n$  e também dos valores dos parâmetros desconhecidos.



**Exemplo.** Vamos considerar os seguintes dados para mostrar o comando que realiza o teste de bondade de ajuste de Anderson-Darling para a distribuição especificada pelo argumento null.

Assume-se que os valores em  $x$  são uma amostra aleatória com alguma função de distribuição  $F$ . A hipótese nula é que  $F$  é a função especificada pelo argumento null, enquanto a hipótese alternativa é que  $F$  seja uma outra função.

```
> library(DescTools)
> AndersonDarlingTest(dados, null = "pnorm", mean=0, sd=1)
```

```
Anderson-Darling test of goodness-of-fit
Null hypothesis: Normal distribution
with parameters mean = 0.000, sd = 1.000
```

```
data: dados
An = 0.50595, p-value = 0.7387
```

O Teorema Central do Limite afirma que, sujeito a que o segundo momento é finito, a média amostral  $\bar{X}$  de observações independentes identicamente distribuídas é assintoticamente normal no sentido de que

$$\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \xrightarrow{D} Z, \quad n \rightarrow \infty,$$

onde  $Z \sim N(0, 1)$ ,  $\mu = E(X_1)$  e  $\sigma^2 = \text{Var}(X_1)$ .

Ao longo dos anos, esse surpreendente resultado informou ou até mesmo confundiu seus usuários. Por exemplo, ele diz que não importa qual seja a distribuição populacional da amostra, a distribuição limite no lado direito é sempre a mesma: a distribuição normal padrão. Imagine quão diferente a distribuição da população pode ser em termos de sua forma: simétrica, distorcida, bimodal, contínua, discreta e assim por diante. No entanto, elas não fazem diferença enquanto ao Teorema do Limite Central.

No entanto, o teorema central do limite é correto do ponto de vista teórico e isso foi confirmado por inúmeros estudos empíricos. Aqui, do ponto de vista teórico, isso significa que  $n \rightarrow \infty$  ou pelo menos é muito grande. No entanto, em uma situação de amostra finita, pode ser uma história diferente.

Por exemplo, suponha que  $n = 30$ . Pode ser mostrado que neste caso a forma da distribuição da população faz diferença. Isso levanta uma questão sobre a taxa de convergência do teorema central do limite. Em particular, duas medidas da forma da distribuição da população são a assimetria e a curtose, definidas como

$$\kappa_3 = \frac{E(X_1 - \mu)^3}{\sigma^3}$$

e

$$\kappa_4 = \frac{E(X_1 - \mu)^4}{\sigma^4} - 3,$$

respectivamente.

Seria de se esperar que essas características tivessem algum impacto na taxa de convergência do teorema central do limite. Por exemplo, o célebre teorema de Berry-Essen, descoberto por Berry (1941) e Essen (1942), afirma que se o terceiro momento de  $X_1$  é limitado, então

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{cE(|X_1|^3)}{\sqrt{n}},$$

onde  $F_n$  é a função de distribuição de  $\xi_n = (\sqrt{n}/\sigma)(\bar{X} - \mu)$ ,  $\Phi$  a função de distribuição normal padrão e  $c$  uma constante que não depende da distribuição de  $X_1$ .

A expansão de Edgeworth, batizada em homenagem ao matemático irlandês Francis Ysidro Edgeworth (1845-1926), é uma forma de aprimorar a desigualdade acima a ordens superiores.

Como a expansão de Taylor, a expansão de Edgeworth pode ser expressa em  $k+1$  termos mais um resíduo. A diferença é que, na expansão de Taylor, os termos estão em ordens decrescentes de  $|x - a|$  e na expansão de Edgeworth, os termos estão em ordem decrescente de  $n^{-1/2}$ . Por uma questão de simplicidade, focamos principalmente no caso  $k = 2$ , que pode ser expresso como

$$F_n(x) = \Phi(x) + \frac{\kappa_3 p_1(x)}{6\sqrt{n}} \phi(x) + \frac{\kappa_4 p_2(x) - \kappa_3^2 p_3(x)}{24n} \phi(x) + O(n^{-3/2}),$$

onde  $\phi(x)$  é a função de densidade normal padrão, ou seja,  $\phi(x) = (1/\sqrt{2\pi})e^{-x^2/2}$ ,

$$p_1(x) = 1 - x^2, \quad p_2(x) = x(3 - x^2) \quad \text{e} \quad p_3(x) = \frac{x}{3}(15 - 10x^2 + x^4).$$

Vemos que o primeiro e o segundo termos da expansão de Edgeworth envolvem a assimetria  $\kappa_3$  e a curtose  $\kappa_4$ , confirmando nossa especulação anterior de que essas quantidades podem influenciar a taxa de convergência do teorema central do limite. Também podemos perceber que se  $\kappa_3 = 0$  e  $\kappa_4 = 0$ , então a aproximação à distribuição normal padrão seria quase perfeita.

Na análise univariada de dados, uma das hipóteses mais utilizadas é a suposição de normalidade. Além disso, a normalidade comumente assumida nos ajuda a estimar e fazer comparações e julgamentos inferenciais. No entanto, a violação dessa suposição pode produzir inferências enganosas e o resultado do uso de inferências não confiáveis é produzir interpretações enganosas. Testes de normalidade devem ser tão importantes quanto a suposição de normalidade.

O método mais amplamente utilizado, pelo menos na econometria, que foi sugerido e usado para testar se a distribuição subjacente a uma amostra é normal é a estatística de Bowman and Shenton (1975):

$$JB = n \left( \frac{\kappa_3^2}{6} + \frac{\kappa_4^2}{24} \right),$$

que posteriormente foi derivado por Jarque and Bera (1987). Por esse motivo, a estatística de teste  $JB$  também chamado de teste de Jarque-Bera.

A estatística  $JB$  tem uma distribuição assintótica chi-quadrada com dois graus de liberdade. Esta estatística é simples e seu poder se mostrou comparável a outros testes poderosos. Está programada na função **JarqueBeraTest** no pacote **DescTools**.

**Exemplo.** Vamos considerar os seguintes dados para mostrar o gráfico p-p plot.

```
> library(DescTools)
> JarqueBeraTest(dados)
```

Robust Jarque Bera Test

```
data: dados
X-squared = 0.45824, df = 2, p-value = 0.7952
```

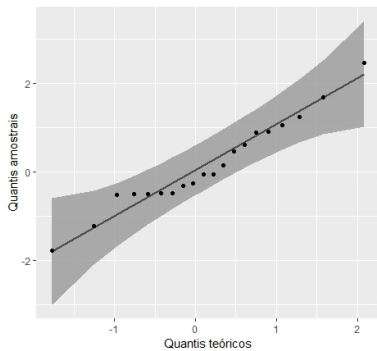
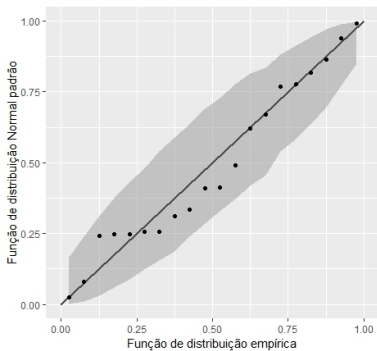


Dois tipos de gráficos são populares na prática. O primeiro é o chamado gráfico de probabilidade, que é na verdade um gráfico de probabilidade versus probabilidade ou p-p plot. Este gráfico também é chamado de plotagem percentual, por razões óbvias. Em termos gerais, o gráfico p-p plot teórico é o gráfico de uma função de distribuição  $F(x)$  versus uma função de distribuição  $G(x)$  para todos os valores de  $x$ .

Como as funções de distribuição são probabilidades, o gráfico p-p plot é convenientemente confinado ao quadrado unitário. Se as duas funções de distribuições forem idênticas o gráfico teórico p-p plot será a diagonal principal, a linha de 45 graus através da origem. Este gráfico é amplamente utilizado para avaliar a assimetria de uma distribuição.

**Exemplo.** Vamos considerar os seguintes dados para mostrar o gráfico p-p plot.

```
> library(qqplotr); library(ggplot2)
> # p-p plot
> gg = ggplot(data = data.frame(dados),
  mapping = aes(sample = dados)) +
  stat_pp_band() +
  stat_pp_line() +
  stat_pp_point() +
  labs(x = "Função de distribuição empírica",
    y = "Função de distribuição Normal padrão")
> gg
```



```
> # q-q plot
> gg1 = ggplot(data = data.frame(dados),
  mapping = aes(sample = dados)) +
  stat_qq_band() +
  stat_qq_line() +
  stat_qq_point() +
  labs(x = "Quantis teóricos", y = "Quantis amostrais")
> gg1
```

Para concluir percebemos que o gráfico p-p plot compara uma amostra com um modelo teórico e representa a proporção teórica menor ou igual a cada valor observado em relação à proporção real.

O segundo tipo de gráfico é o chamado quantile plot, que é na verdade um gráfico quantil versus quantil ou um gráfico q-q plot. O gráfico q-q teórico é o gráfico dos quantis de uma função de distribuição  $F$  versus os correspondentes quantis da função de distribuição  $G$ , isto é, mostramos os pontos  $(F^{-1}(p), G^{-1}(p))$  para  $0 < p < 1$ . Se as duas funções de distribuição forem idênticas, o gráfico q-q teórico será a diagonal principal, a linha de 45 graus através da origem. Se  $F(x) = G\left(\frac{x-\mu}{\sigma}\right)$ , é facilmente visto que  $F^{-1}(p) = \mu + \sigma G^{-1}(p)$ , de modo que o p-ésimo quantil de  $F$  e  $G$  tem uma relação linear. Assim, se duas distribuições diferirem apenas em localização e/ou escala, o gráfico teórico q-q será uma reta com inclinação  $\sigma$  e intercepto  $\mu$ .