Verificando a normalidade

Gilberto Pereira Sassi

Universidade Federal da Bahia
Instituto de Matemática e Estatística
Departamento de Estatística

Objetivo

- Checar se uma variável aleatória contínua X tem modelo de probabilidade normal;
- Vamos usar:
  - histograma;
  - gráfico de quantis;
  - teste hipóteses de Kolmogorov-Smirnov;
- Lembre que só podemos usar o modelo de probabilidade t-Student para o intervalo de confiança e o teste de hipóteses para uma variável aleatória contínua com modelo de probabilidade normal e variância populacional desconhecida.

Histograma

Exemplo

Devido ao surgimento de um novo vírus, uma empresa começou a vender gel anti-séptico para as mãos. Os frascos conteriam 60*ml* do produto, mas pequenas variações são comuns devido a fatores incontroláveis na produção. Para estudar se a fabricação cumpre os critérios pré-estabelecidos, 20 frascos foram selecionados aleatoriamente em um certo dia, com os seguintes resultados:

frascos	1	2	3	4	5	6	7	8	9	10
conteúdo	60,8	61,5	60,5	60,9	59,9	60,2	63,9	59,5	59,7	62,5
frascos	11	12	13	14	15	16	17	18	19	20
conteúdo	59,9	60,5	60,7	55,6	57,8	58,0	57,8	61,6	59,3	62,6

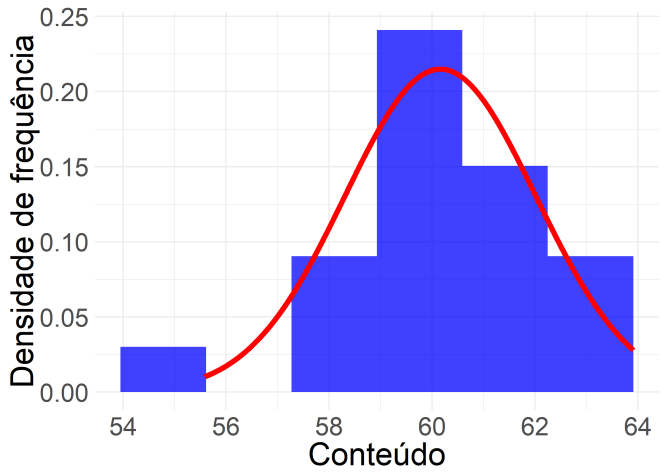
Verifique se a variável `conteúdo` tem distribuição normal.

Histograma

Solução

- A ideia é desenhar o histograma e desenhamos a função de densidade da variável aleatória $N(\bar{x}, s_x)$: se o formatado do histograma e a função de densidade foram iguais, temos indícios de normalidade dos dados;
- Para determinar o número de classes, você pode usar a regra de Sturge: Número de faixas $= 1 + \lceil \log_2(n) \rceil$, em que n é o tamanho da amostra.

Figura 1: Histograma e a função de densidade da distribuição normal.



Pela Figura 1, existe indício de normalidade dos dados.

Gráfico de quantis para duas variáveis

Gráfico de quantis ou Q-Q plot

- Usamos para verificar se duas variáveis aleatórias têm a mesma distribuição;
- Imagine que as amostras de X e Y tem o mesmo tamanho e considere as estatísticas de ordem de X e Y:
  - $x_{(j)}$: j-ésimo menor valor da amostra de X;
  - $y_{(j)}$: j-ésimo menor valor da amostra de Y;

  Então cada par $(x_{(j)}, y_{(j)})$, $j = 1, \ldots, n$ é representado por um ponto no plano cartesiano;
- Seja X e Y duas variáveis quantitativas com valores observados:

$$\begin{aligned} X: &\quad x_1, \ldots, x_n, \\ Y: &\quad y_1, \ldots, y_m, \end{aligned}$$

  em que $m < n$.

  Então cada par $(q_{\left(\frac{j}{m}\right)}, x_{(j)})$, $j = 1, \ldots, m$, em que $q_{\left(\frac{j}{m}\right)}$ é o quantil de ordem $\frac{j}{m}$ da variável X;
- Se os pontos estiverem próximos da reta de 45°, temos uma indicação de que as duas variáveis tem o mesmo modelo de probabilidade.

Gráfico de quantis para duas variáveis

Exemplo

Suponha que desejamos comparar as alturas (em metros) de alunos de ensino médio de duas escolas *A* e *B*. Uma amostra de 15 estudantes, foi sorteada em cada um dessas e os resultados são apresentados a seguir:

Tabela 1: Alturas (em *cm*) para as escolas *A* e *B*.

Escola *A*	146,5	162,2	167,1	161,6	147,9
	155,6	167,4	178,8	156,1	144,1
	165,9	165,6	164,1	161,9	169,2
Escola *B*	160,4	153,1	156,3	156,8	175,4
	154,1	155,4	160,3	163,4	174,7
	161,9	151,8	165,0	161,7	138,0

Gráfico de quantis para duas variáveis

Solução

Valores ordenados das alturas para as duas escolas *A* e *B*:

Tabela 2: Alturas (em *cm*) para as escolas *A* e *B*.

Escola *A*	144,1	146,5	147,9	155,6	156,1
	161,6	161,9	162,2	164,1	165,6
	165,9	167,1	167,4	169,2	178,8
Escola *B*	138,0	151,8	153,1	154,1	155,4
	156,3	156,8	160,3	160,4	161,7
	161,9	163,4	165,0	174,7	175,4

Gráfico de quantis para duas variáveis quantitativas

Solução

No gráfico da Figura 2, mostramos o gráfico de quantis para as duas variáveis. A linha azul é a reta com inclinação de 45° e os pontos estão próximos e em torno desta reta. Logo, as duas variáveis tem o mesmo modelo de probabilidade.

Figura 2: Gráfico de quantis para as alturas das crianças na Escola A e na Escola B.

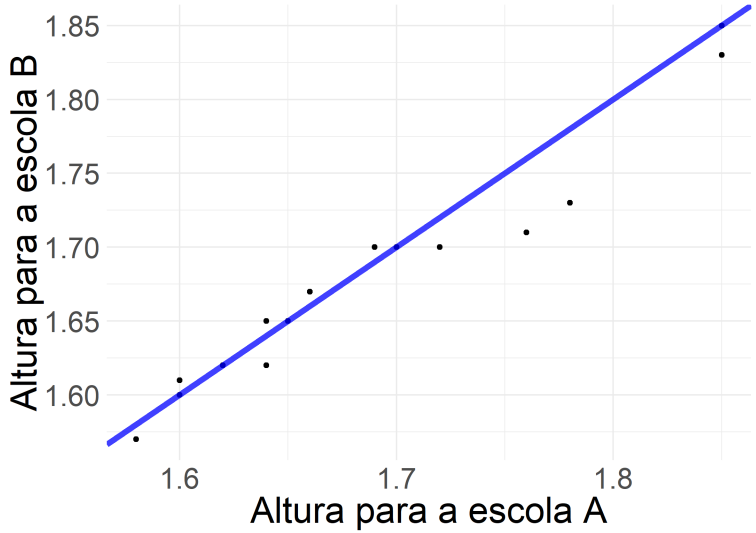


Gráfico de probabilidade normal

Imagine que

- X com valores observados $x_1, \ldots, x_n$ e estatística de ordem $x_{(1)}, \ldots, x_{(n)}$;
- Média populacional: μ e desvio padrão populacional: σ. Na ausência de μ e σ, podemos usar $\bar{x}$ e s.

Considere

$$z_{(i)} = \frac{x_{(i)} - \mu}{\sigma}, \ i = 1, \ldots, n,$$
$$\Phi\left(q_{(i)}\right) = \frac{i - 0,5}{n}, \ i = 1, \ldots, n.$$

em que $P(Z < z) = \Phi(z)$ e $Z \sim N(0, 1)$.

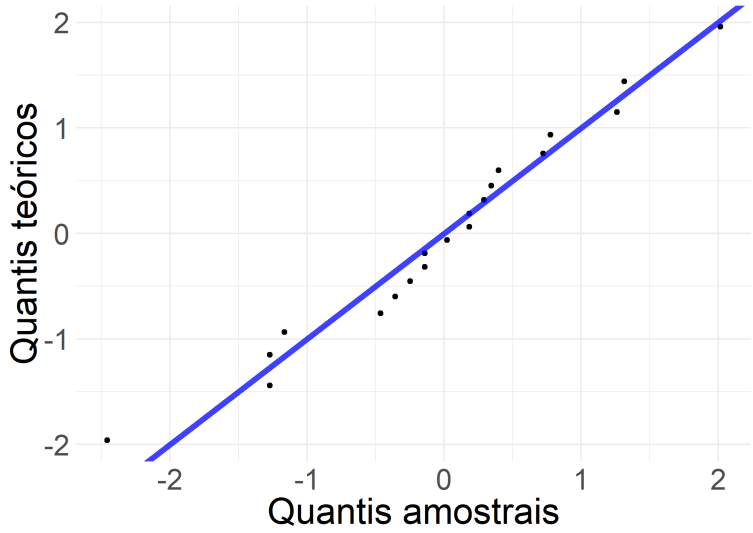
Para cada par $(z_{(i)}, q_{(i)}), \ i = 1, \ldots, n$, desenho no ponto um plano cartesiano. Chamamos este gráfico de quantis de gráfico de probabilidade normal.

Gráfico de probabilidade normal

Solução (continuação)

No gráfico da Figura 3, mostramos o gráfico de quantis de probabilidade normal para a variável `conteúdo`. A linha azul é a reta com inclinação de 45° passando pela origem e os pontos estão próximos e em torno desta reta. Logo, temos indício de que a variável `conteúdo` tem distribuição normal.

Figura 3: Gráfico de quantis: distribuição normal.



Teste de Kolmogorov-Smirnov

Objetivo

Checar se uma variável X tem um determinado modelo de probabilidade. De forma similar, queremos checar se a função de distribuição acumulada $F_X(x)$ da variável é $F_0(x)$, então temos as hipóteses:

$$H_0 : F_X(x) = F_0(x), \forall x,$$
$$H_1 : F_X(x) \neq F_0(x), \text{para algum valor de } X.$$

Teste de Kolmogorov-Smirnov

Exemplo

Considere uma amostra de 15 equipamentos que foram testados até a sua falha. Verifique se essa variável tem distribuição exponencial com média $\mu = 8000$ horas usando o teste de Kolmogorov-Smirnov ao nível de significância $\alpha = 5\%$.

785,17	1002,96	1373,79	1487,74	2557,56
4348,16	7232,38	9491,11	9760,40	12878,08
15047,12	15622,29	18460,61	20023,98	26195,43

Tabela 3: Tempo de vida de um equipamento.

Teste de Kolmogorov-Smirnov

Solução

Passo 1) Primeiro vamos estabelecer as hipóteses:

$$H_0 : F_X(x) = F_0(x),$$
$$H_1 : F_X(x) \neq F_0(x),$$

em que $F_0(x)$ é a função de distribuição acumulada do modelo exponencial com média $\mu = 8000$ horas.
Passo 2) Nível de significância $\alpha = 5\%$. (Tabela tem valores apenas para 5%, 2% e 1%).
Passo 3) Rejeitamos H_0 se D for grande. Ou seja, $RC = \{D \mid D \geq x_c\}$, em que

$$D = \max_{1 \leq i \leq n} |F(x_i) - F_e(x_i)| = \max_{1 \leq i \leq n} |F(x_{(i)}) - F_e(x_{(i)})| = \max_{1 \leq i \leq n} \left| F(x_{(i)}) - \frac{i}{n} \right|.$$

e $F_e(x) = \frac{N(x)}{n}$ é uma aproximação da função de distribuição acumulada da variável aleatória X.

Teste de Kolmogorov-Smirnov

Solução

Passo 4) Encontrando o nível de significância: $\alpha = 5\%$.

$$\begin{aligned}\alpha &= P(\text{Decidir por } H_1 \mid H_0 \text{ é verdadeiro}) \\ &= P(D > x_c \mid H_0 \text{ é verdadeiro}).\end{aligned}$$

Para encontrar x_c, usamos a tabela X do livro `Estatística Básica`, e $x_c = 0,264$.
Passo 5) Considere $F(x) = 1 - \exp\left(-\frac{x}{\mu}\right),\ x \geq 0$. Na tabela 4, calculamos $\left|F(x_{(i)}) - \frac{i}{n}\right|, i = 1, \ldots, n$, e notamos que $D = \max_{1 \leq i \leq n} \left|F(x_{(i)}) - \frac{i}{n}\right| = 0,1613$.

i	$x_{(i)}$	$F(x_{(i)}) = 1 - \exp\left(-\frac{x_{(i)}}{\mu}\right)$	$\frac{i}{n}$	$\left\|1 - \exp\left(-\frac{x_{(i)}}{\mu}\right) - \frac{i}{n}\right\|$
1	785,17	0,0935	0,0667	0,0268
2	1002,96	0,1178	0,1333	0,0155
3	1373,79	0,1578	0,2000	0,0422
4	1487,74	0,1697	0,2667	0,0970
5	2557,56	0,2736	0,3333	0,0597
6	4348,16	0,4193	0,4000	0,0193
7	7232,38	0,5951	0,4667	0,1284
8	9491,11	0,6947	0,5333	0,1613
9	9760,40	0,7048	0,6000	0,1048
10	12878,08	0,8001	0,6667	0,1334
11	15047,12	0,8475	0,7333	0,1142
12	15622,29	0,8581	0,8000	0,0581
13	18460,61	0,9005	0,8667	0,0338
14	20023,98	0,9182	0,9333	0,0152
15	26195,43	0,9622	1,0000	0,0378

Tabela 4: Calculando o valor de D.

Como $D = 0,1613 \leq 0,338$ e decidimos por H_0, o seja, ao nível de significância 5% decidimos que a variável X tem distribuição exponencial com média $\mu = 8000$.

Teste de Kolmogorov-Smirnov

Solução

Vamos voltar ao exemplo da máquina que enche potes com álcool gel.
Passo 1) Queremos verificar as seguintes hipóteses:

$$H_0 : \texttt{conteúdo} \text{ tem distribuição normal;}$$
$$H_1 : \texttt{conteúdo} \text{ não tem distribuição normal;}$$

Passo 2) Nível de significância $\alpha = 5\%$.
Passo 3) Rejeitamos H_0 se D for grande. Ou seja, $RC = \{D \mid D > x_c\}$, em que $D = \max_{1 \leq i \leq n} \left| \Phi\left(\frac{x_{(i)} - \bar{x}}{dp(X)}\right) - \frac{i}{n} \right|$. Ou seja, a região crítica é $RC = \{D \mid D > x_c\}$.
Passo 4) Usando a tabela X, temos que o valor crítico é $x_c = 0,294$.

Verificando a normalidade dos dados

Passo 5) Usando a Tabela 5, percebemos que $D = 0,0950 \leq 0,294 = D$ e não rejeitamos H_0. Ou seja, ao nível de significância $\alpha = 5\%$, a variável aleatória `conteúdo` tem distribuição normal.

i	$x_{(i)}$	$z_{(i)} = \frac{x_{(i)} - \bar{x}}{dp(x)}$	$\Phi(z_{(i)})$	$\frac{i}{n}$	$\lvert\Phi(z_{(i)}) - \frac{i}{n}\rvert$
1	55,6	-2,4570	0,0070	0,0500	0,0430
2	57,8	-1,2716	0,1018	0,1000	0,0018
3	57,8	-1,2716	0,1018	0,1500	0,0482
4	58,0	-1,1638	0,1222	0,2000	0,0778
5	59,3	-0,4634	0,3215	0,2500	0,0715
6	59,5	-0,3556	0,3611	0,3000	0,0611
7	59,7	-0,2479	0,4021	0,3500	0,0521
8	59,9	-0,1401	0,4443	0,4000	0,0443
9	59,9	-0,1401	0,4443	0,4500	0,0057
10	60,2	0,0216	0,5086	0,5000	0,0086
11	60,5	0,1832	0,5727	0,5500	0,0227
12	60,5	0,1832	0,5727	0,6000	0,0273
13	60,7	0,2910	0,6145	0,6500	0,0355
14	60,8	0,3448	0,6349	0,7000	0,0651
15	60,9	0,3987	0,6550	0,7500	0,0950
16	61,5	0,7220	0,7649	0,8000	0,0351
17	61,6	0,7759	0,7811	0,8500	0,0689
18	62,5	1,2608	0,8963	0,9000	0,0037
19	62,6	1,3147	0,9057	0,9500	0,0443
20	63,9	2,0152	0,9781	1,0000	0,0219

Tabela 5: Normalidade dos dados.