Introdução à datascience com R

Cleuton Sampaio

Lição 4: Probabilidades

Probabilidades

Probabilidade é um valor contínuo que nos diz a chance de ocorrência de determinado resultado entre vários outros, de um conjunto de resultados possíveis. Vamos recorrer a um dado (sim, desses que você usa em jogos). Um dado tem 6 faces, numeradas de 1 a 6. Qual é a probabilidade de você lançar um dado e sair um número 4?

- Número de faces: 6;
- Faces que nos interessam: 1 (aquela que contém o número 4);
- Probabilidade de sair um 4: 1/6 ou aproximadamente 17%.

Por que probabilidade?

Existe um ramo da estatística, chamado "Inferência estatística", que se preocupa em fazer afirmações sobre as variáveis de uma população, com base em amostras e, para isto, utiliza entre outras ferramentas o estudo da distribuição de probabilidades dos seus valores.

Segundo a Wikipedia:

"A inferência estatística faz proposições sobre um universo, usando dados tirados de um universo com algum tipo de amostragem. Dada um hipótese sobre um universo, para o qual nós queremos tirar inferências, a inferência estatística consiste em (primeiramente) selecionar um modelo estatístico do processo que gera os dados e (segundamente) deduzir as proposições a partir do modelo.
Konishi & Kitagawa afirmam, "A maior parte dos problemas na inferência estatística podem ser considerados problemas relacionados à modelagem estatística". De forma relacionada, Sir David Cox disse que, "Como a tradução do problema da matéria é feita para o modelo estatístico é com frequência a parte mais crítica de uma análise".
A conclusão de uma inferência estatística é uma proposição estatística. Algumas formas comuns de proposições estatísticas são as seguintes:

- Estimativa por ponto. Ex. Um valor particular que melhor aproxima algum parâmetro de interesse;
- Estimativa por intervalo. Ex. Um intervalo de confiança (ou estimativa por conjunto), ex. um intervalo construído ao usar um conjunto de dados tirados de um universo de forma que, baixo amostragens repetidas de tais conjuntos de dados, tais intervalos conteriam o verdadeiro valor parâmetro com a probabilidade no dito nível de confiança.
- Intervalo de credibilidade. Ex. um conjunto de valores contendo, por exemplo, 95% de crença posterior.
- Rejeição de uma hipótese.
- Clustering ou classificação de pontos de dados em grupos."

Para fazer inferências estatísticas, basicamente temos que supor a geração e a distribuição dos valores das variáveis observadas. Para isto, usamos modelos estatísticos e assumimos determinados comportamentos com base em probabilidades.

Vamos supor que você seja o responsável pelo controle de qualidade de um laboratório farmacêutico, e tenha recebido reclamações que em alguns lotes de determinado medicamento, a quantidade de um composto era menor que o esperado. Como lidaria com isso? Certamente, colheria amostras e verificaria se a reclamação tem fundamento, ou seja, buscaria evidências estatísticas que confirmem ou rejeitem a hipótese de que todos os lotes contém a mesma quantidade do composto. Em outras palavras, você analisaria a probabilidade, com evidências estatísticas, da ocorrência destas anomalias.

Variável aleatória

Uma variável aleatória possui um valor único para cada resultado de experimento ou observação. Vamos supor que estejamos estudando o peso dos alunos de uma turma. Podemos selecionar um aluno aleatoriamente e verificar seu peso.

Uma variável pode ser contínua ou discreta, conforme já vimos anteriormente.

Selecionando um aluno da turma, qual a probabilidade de seu peso ser próximo da média? Para estudar a distribuição dos pesos, e fazer predições, podemos criar uma **distribuição de probabilidades** com nossas variáveis aleatórias.

Valor esperado

Você verá muito este termo. Quando estamos estudando um fenômeno, expresso através de uma variável aleatória, estamos interessados no seu valor central, ou sua média. Chamamos essa média de "valor esperado" ou "esperança" da variável aleatória. Por exemplo:

- Você está esperando um determinado ônibus. Olha para o relógio, vira-se para a pessoa ao lado e pergunta: “Quanto tempo este ônibus demora para passar aqui?” A variável aleatória é o tempo de espera (ou o período de espera). A pessoa responde: “Passa de 15 em 15 minutos.” Então, sabendo que o último ônibus acabou de passar, você se concentra no valor esperado de 15 minutos;
- Um cliente típico da nossa App móvel adquire um produto a cada 50 acessos. Logo, você espera que, ao chegar perto de 50 acessos, os clientes adquiram produtos.

Frequências

Frequentemente, dividimos os valores em faixas ou classes, para facilitar nossa visualização. Isto é especialmente interessante quando temos dados contínuos (valores reais) ou categorias. Por exemplo, supondo que temos um dataset com os gastos mensais das famílias Brasileiras, como poderíamos analisá-lo? O valor do gasto é contínuo, logo, não adianta querer contar quantas famílias têm o mesmo gasto. Melhor seria criar classes de gastos, certo?

Se separarmos a faixa total de valores de gastos em classes, podemos contar quantas vezes uma família se encaixa naquela classe de gastos. Isto se chama frequência.

Histogramas

Um histograma é um gráfico das frequências dos elementos de um dataset. É claro que existem definições mais complexas do que esta, mas é a mais simples possível.

Um histograma é uma distribuição de frequências e podemos fazer estudos interessantes sobre ela.

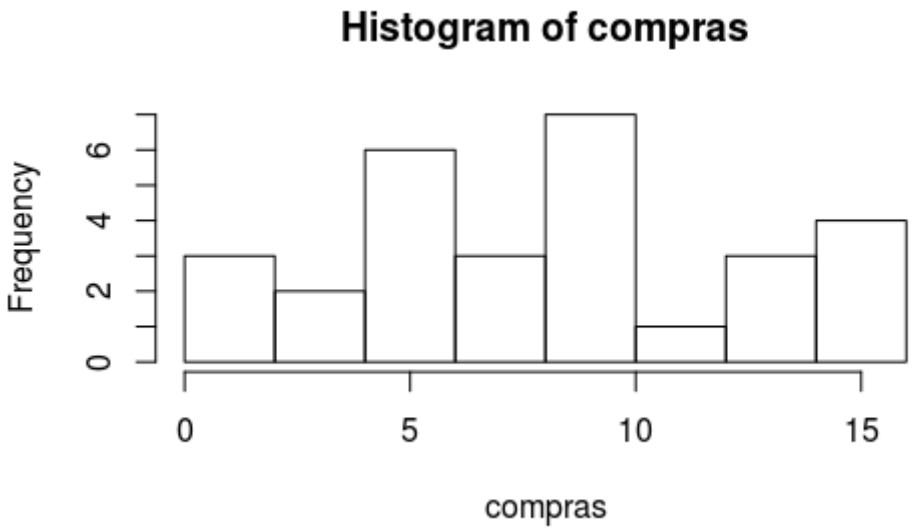
Criar um histograma é uma verdadeira arte… Vamos supor o seguinte conjunto de dados, que corresponde ao total de itens comprados em determinado dia, em cada pedido:

```
compras <-
c(1,1,1,3,3,5,5,6,6,6,6,7,8,8,9,9,9,9,10,10,10,11,13,14,14,15,15,15,15)
```

Temos as medidas básicas desse conjunto:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   6.000   9.000   8.414  11.000  15.000
```

Se tentarmos criar um histograma com a função padrão ("hist()") veremos algo assim:



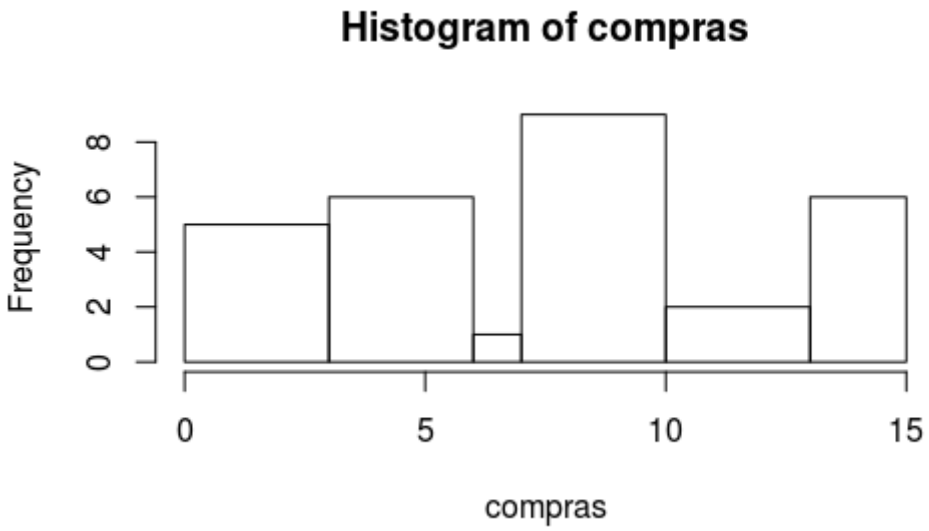
A divisão dos valores em classes não foi a ideal, gerando alternância entre alto e baixo. A técnica para construir histogramas úteis têm alguns princípios:

1. As classes de valores deverão ter amplitudes iguais;
2. O número de intervalos não deve ultrapassar 20;
3. Escolher limites que facilitem o agrupamento;
4. Ao construir o histograma, cada retângulo deverá ter área proporcional à frequência relativa;
5. Calcular a quantidade de intervalos de classes. O método de Sturges usa uma fórmula: k = (3,3 * log n + 1).

A Wikipedia tem um guia bem interessante: https://pt.wikipedia.org/wiki/Histograma

Eu calculei manualmente a quantidade de classes e escolhi os valores limítrofes de cada intervalo, passando isso para a função "hist()" no barâmetro "breaks":

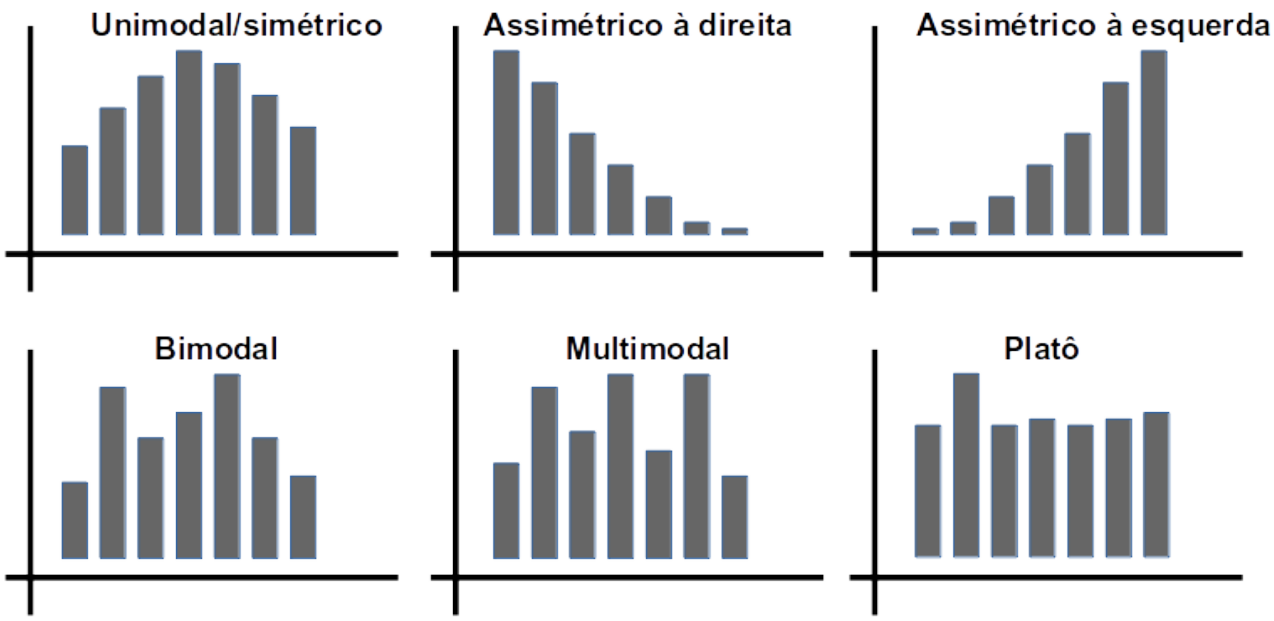
```
print(hist(compras, breaks = c(0,3,6,7,10,13,15), freq = TRUE))
```



Ficou um pouco melhor e bem aproximado do que eu calculei manualmente. O que notamos logo nesse gráfico? Que ele parece ter uma forma de sino ou de montanha, exceto pelas duas classes "(6,7]" e "(10,13]" que estão destoando. Não fosse por isto, ele teria uma curva suave que sobe até a classe "(7,10]" e desce até a classe final.

Interpretando histogramas

A forma da distribuição dos valores em torno da média é nos dá a assimetria da distribuição:



Podemos medir a assimetria de uma distribuição com a função "skewness()":

```
print(paste('Assimetria:',skewness(compras)))
```

```
"Assimetria: -0.0256677795203193"
```

Valores negativos indicam que a cauda "esquerda" do sino é maior que a da direita, e vice-versa. Valor nulo indica que a distribuição é perfeitamente simétrica.

Nosso conjunto de dados tem mais valores abaixo da média.

Distribuição de probabilidades

Além de estudar o passado, a estatística também nos auxilia a estudar o futuro, ou as probabilidades de ocorrências de eventos, através da construção de modelos preditivos.

Por exemplo, estudando as variáveis estatísticas de um dataset, como os pedidos dos clientes de uma loja virtual, é possível deduzir seus hábitos de consumo e até criar um modelo que nos permita predizer quanto será vendido e de quais itens.

Para isto, temos que falar um pouco sobre probabilidades.

Podemos atribuir probabilidades a cada valor possível de uma variável aleatória, criando, desta forma uma distribuição de probabilidades de cada valor ocorrer.

Se estivermos falando de variáveis discretas, usaremos uma **Função massa de probabilidades** (Probability Mass Function) para associar cada valor possível a uma probabilidade. Isto é conhecido como PMF.

Se estivermos falando de variáveis contínuas, usaremos uma **Função de densidade de probabilidades** (Probability Density Function) que diz a probabilidade relativa da variável assumir um valor dado. Isto é conhecido como PDF.

Modelos probabilísticos discretos

Se estamos estudando um fenômeno, cuja variável aleatória é discreta, então temos alguns modelos de distribuições de probabilidade, que nos ajudam a entender e criar

modelos preditivos. Vamos ver dois modelos de eventos probabilísticos muito comuns: Binomial e Poisson.

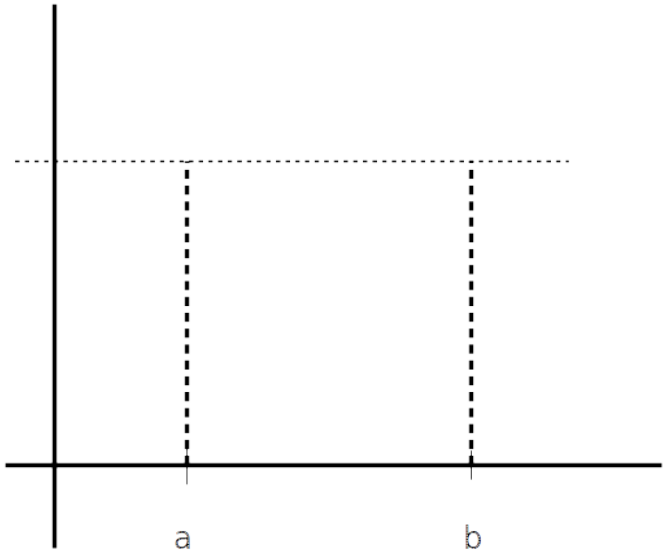
Modelos probabilísticos contínuos

Bem, neste caso, a variável aleatória do fenômeno que estamos estudando é contínua, como peso, velocidade ou custo.

Distribuição uniforme

A distribuição uniforme é muito importante para o estudo de fenômenos, e pode ser entendida como um experimento com um número finito de resultados, todos com chances iguais de acontecerem.

Vamos supor um servidor que tenha probabilidade uniforme de dar pane em 30 dias de uso contínuo. Qual a probabilidade dele dar pane em uma semana de uso?



Distribuição normal ou Gaussiana

Certamente, você já viu um gráfico parecido com esse:

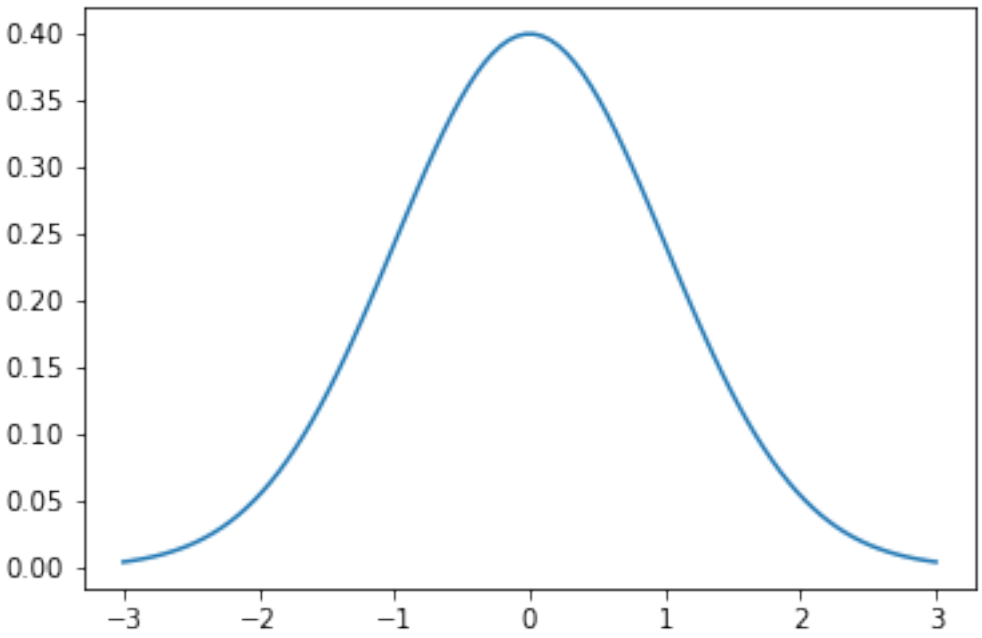
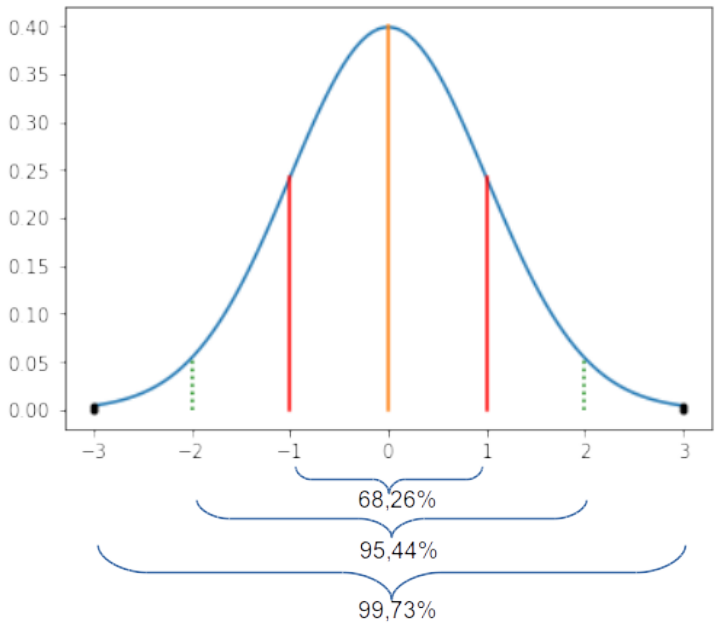


Figura 1: Distribuição normal com média zero

Esta distribuição é importante porque muitos fenômenos **naturais** seguem esse mesmo modelo. Na figura, vemos um gráfico de distribuição normal padronizada, com μ = 0 e σ = 1. Para obtermos esse gráfico a partir de uma variável contínua aleatória *X*, precisamos usar a fórmula:

$$Z=(\frac{(x-\mu)}{\sigma})$$

Existem algumas áreas na distribuição normal que caracterizam a nossa população:



As áreas possuem concentrações de elementos:

- 68,26% dos elementos encontram-se na área com até 1 desvio padrão da média;
- 95,44% dos elementos encontram-se na área com até 2 desvios padrões da média;
- 99,73% dos elementos encontram-se na área com até 3 desvios padrões da média.

A importância da distribuição normal (ou Gaussiana) é dada pelo **Teorema Central do Limite**. Este teorema indica que, quando se aumenta o tamanho da amostra, a distribuição da média dos valores se aproxima da distribuição normal, mesmo que, originalmente, a distribuição da população de ocorrências não siga a distribuição normal.

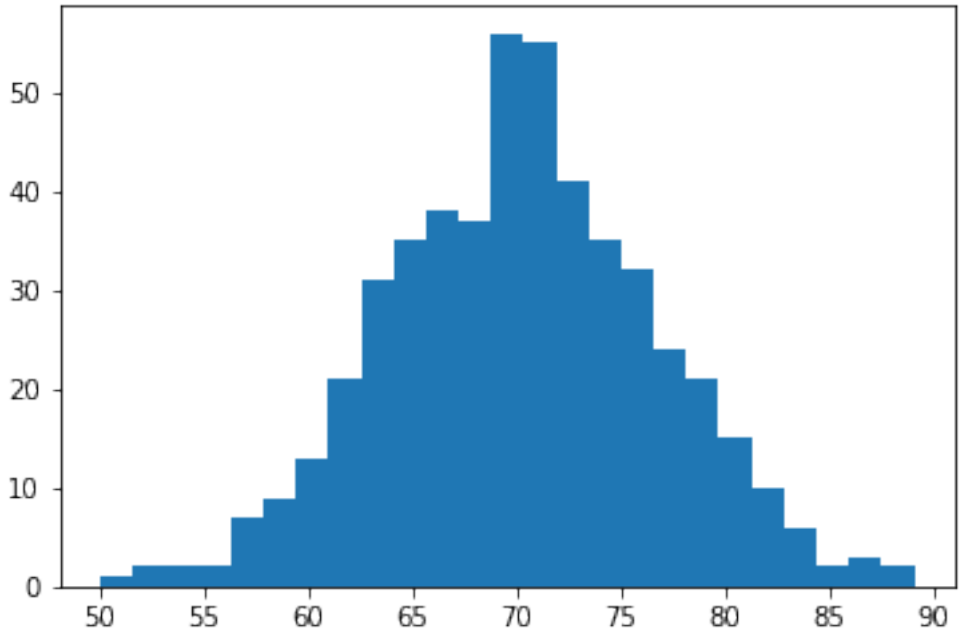
Atenção: Normalmente assumimos a distribuição normal quando a variância da população (σ^2) é conhecida. Caso contrário, assumimos a distribuição T de Student (veremos mais adiante).

A função de densidade de probabilidade (PDF) da distribuição normal é:

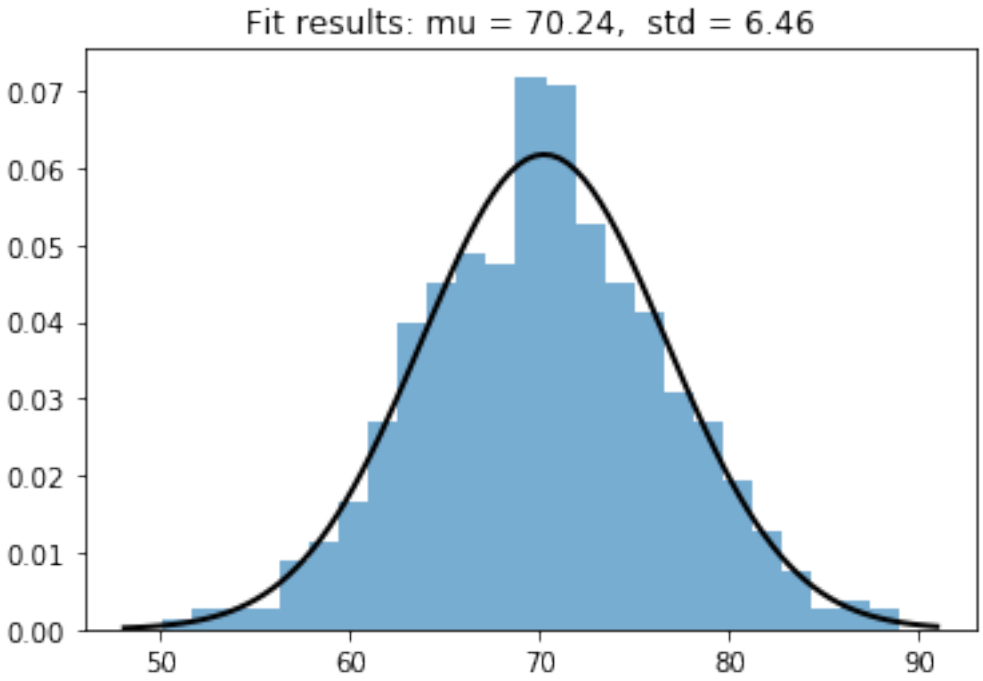
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Vamos dar um exemplo:

Em um grupo de 500 alunos adolescentes de uma escola, foram colhidos seus pesos. Plotamos um histograma com esses pesos:



Sabemos que a média é de 70 kg, e o desvio padrão é 6,5. Agora, podemos notar que esta distribuição se aproxima da distribuição normal:



Vemos que o histograma dos pesos se aproxima da distribuição normal, logo, podemos concluir que a maioria dos alunos pesa entre 70 +- 6,5 Kg, ou, entre 63,5 kg e 76,5 kg.
Existem muito mais modelos de distribuições de variáveis contínuas, como:

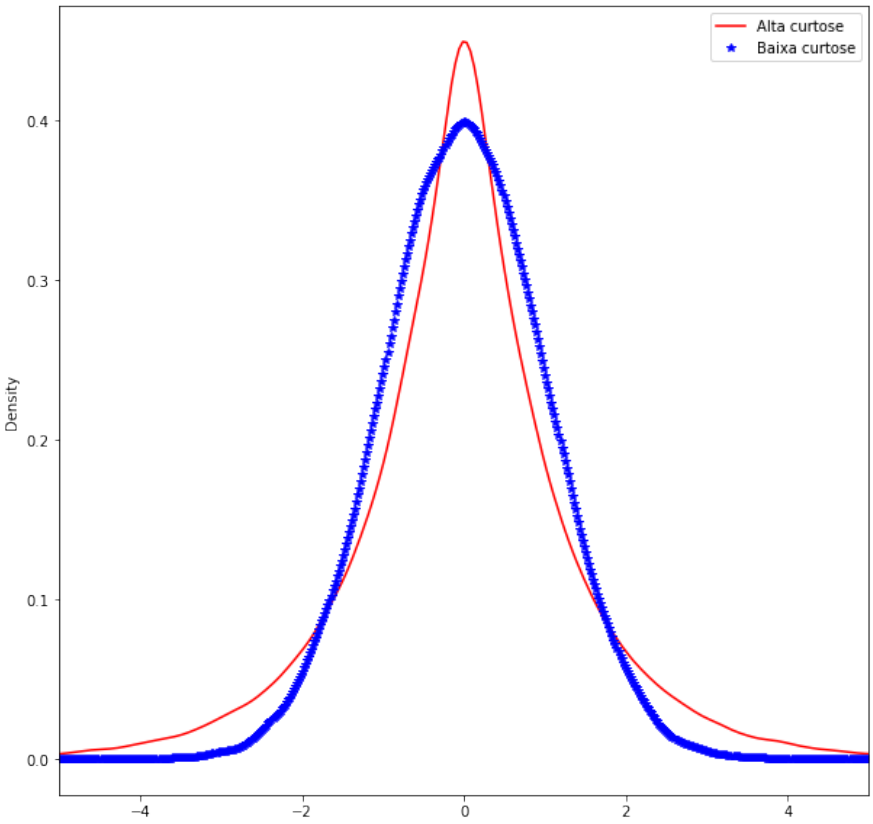
- Qui-quadrado;
- T de Sudent;
- Gama;
- Beta:
- Log-normal;
- Logística;

Curtose e assimetria

Quando temos uma distribuição de probabilidades ou mesmo um dataset com amostras, existem duas medidas que são muito interessantes:

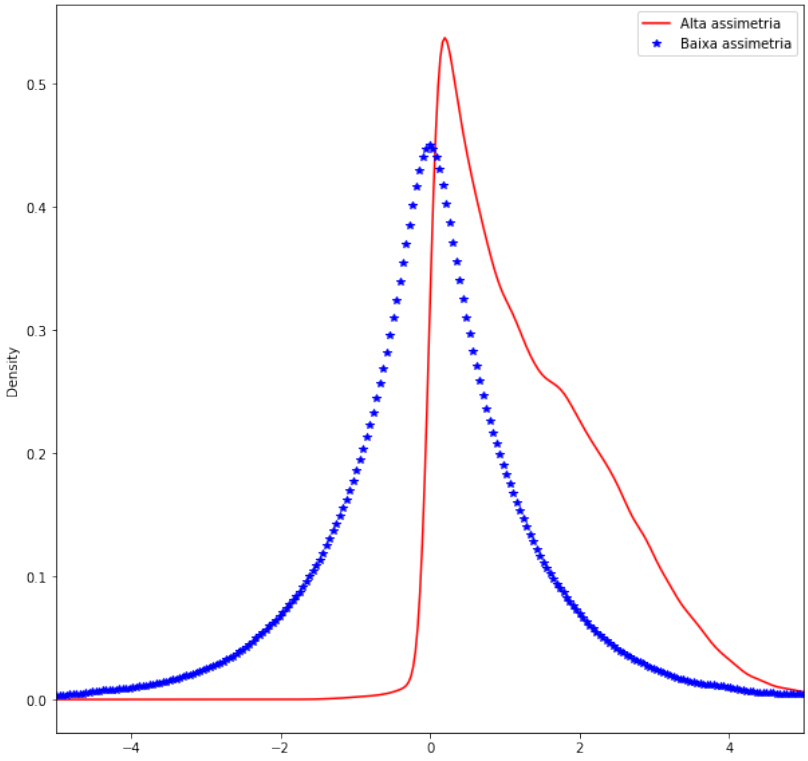
- Curtose (**Kurtosis**): O quão "pontuda" ou "achatada" é a distribuição;
- Assimetria (**Skewness**): A medida da simetria da distribuição em torno da média.

É mais fácil ver do que falar…



Com relação à curtose, uma distribuição pode ser:

- Curtose próxima de zero: Mesocúrtica;
- Curtose maior que zero: Leptocúrtica (pontuda);
- Curtose menor que zero: Platicúrtica (achatada).



Já a assimetria ocorre quando a distribuição não é centrada em torno da média, como vemos na figura. A linha contínua uma "cauda" mais comprida à direita do que à esquerda. Podemos avaliar a assimetria assim:

- Assimetria > 0: Mais valores acima da média à direita (cauda maior à direita);
- Assimetria < 0: Mais valores acima da média à esquerda (cauda maior à esquerda);

O que significa a assimetria **positiva** (direita)? Que há mais valores mais altos que a média (a cauda). Já a assimetria **negativa** significa que há mais valores menores que a média.

Agora em R

Este é um resumo do que vimos no script desta aula. Ele se chama "probabilidades.R" e está no repositório do curso, em:

https://github.com/cleuton/datascience/tree/master/R-course/lession4

Estatística básica

```
# Histograma:
# Itens comprados em nossa loja na última semana, por pedido:
compras <-
c(1,1,1,3,3,5,5,6,6,6,6,7,8,8,9,9,9,9,10,10,10,11,13,14,14,15,15,15,15)
print(summary(compras))
print(paste('Desvio amostral:',sd(compras)))
```

Geramos um vetor das compras utilizando a atribuição "<-" e a função "combine". Depois, usamos a função "summary()" para gerar as medidas de tendência central:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   6.000   9.000   8.414  11.000  15.000
```

- Min: Valor mínimo da amostra;
- 1st Qu.: Valor do primeiro quartil;
- Median: Mediana;
- Mean: Média;
- 3rd Qu.: Terceiro quartil;
- Max: valor máximo.

Para entender o que é um "quartil", há este excelente artigo da Wikipedia:

"Na estatística descritiva, um quartil é qualquer um dos três valores que divide o conjunto ordenado de dados em quatro partes iguais, e assim cada parte representa 1/4 da amostra ou população.

Assim, no caso duma amostra ordenada,

- primeiro quartil (designado por Q1/4) = quartil inferior = é o valor aos 25% da amostra ordenada = 25º percentil;
- segundo quartil (designado por Q2/4) = mediana = é o valor até ao qual se encontra 50% da amostra ordenada = 50º percentil, ou 5º decil;
- terceiro quartil (designado por Q3/4) = quartil superior = valor a partir do qual se encontram 25% dos valores mais elevados = valor aos 75% da amostra ordenada = 75º percentil;

à diferença entre os quartis superior e inferior chama-se amplitude inter-quartil."

Histogramas

```
# Histograma padrão:
print(hist(compras))

# Histograma com classes aproximadamente de igual amplitude:
print(hist(compras, breaks = c(0,3,6,7,10,13,15), freq = TRUE))
```

No histograma padrão, as classes são calculadas automaticamente, e a figura pode ficar comprometida. Podemos calcular as classes e usar o parâmetro "breaks" para informar um vetor, com as medidas que são os limites das classes. Neste caso, temos que informar também o parâmetro "freq", que é binário, como TRUE (verdadeiro).

Assimetria e curtose

Para calcular a assimetria (skewness) e curtose (kurtosis) usamos funções do pacote "moments", e é preciso instalá-lo. Digite na janela "console", na parte inferior esquerda do RStudio:

```
install.packages('moments')
```

Para utilizar funções de pacotes de terceiros, precisamos usar o comando "library()" dentro do nosso código. Então, o cálculo de assimetria e curtose fica assim:

```
library(moments)
print(paste('Assimetria:',skewness(compras)))

# Curtose (Kurtosis):
print(paste('Curtose:',kurtosis(compras)))
```

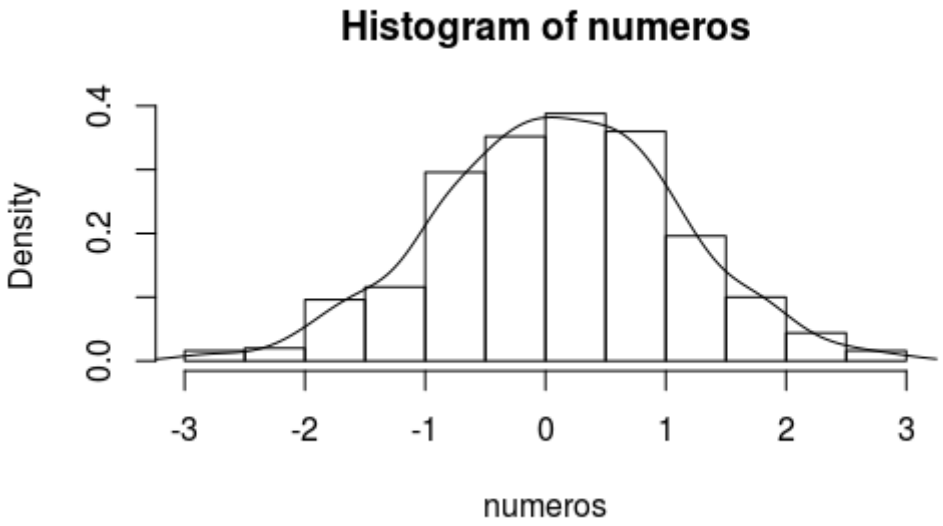
Distribuição normal de probabilidade e números aleatórios

Podemos gerar variáveis aleatórias, obtendo números aleatórios a partir de uma distribuição normal. Para isto, usamos a função "rnorm()":

```
numeros <- rnorm(500)
hist(numeros,probability = TRUE)
print(lines(density(numeros)))
```

Geramos 500 números, selecionados a partir de uma distribuição normal de probabilidades, com média zero e desvio padrão 1 (isso pode ser alterado).

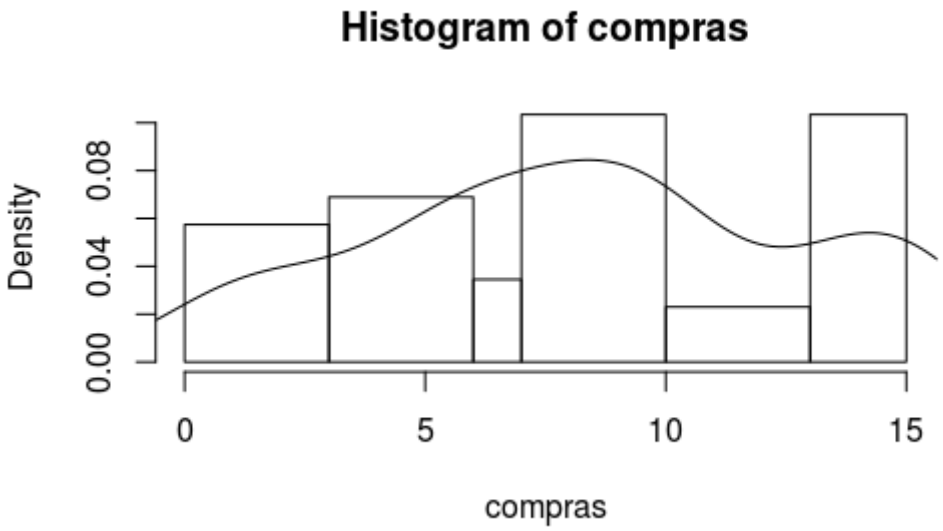
Eis o histograma:



Notou que este histograma apresenta uma "smooth" curve, ou uma curva de tendência? Eu mandei plotar um histograma com densidades em vez de contagem nas ordenadas (PDF). E usei o comando "lines()" para plotar a densidade dos valores.

Podemos fazer isso com qualquer histograma, inclusive o das compras:

```
# Plotando a densidade das compras:
hist(compras, breaks = c(0,3,6,7,10,13,15), probability = TRUE)
print(lines(density(compras)))
```



Isto é muito legal pois nos dá a chance de visualizar melhor a tendência da distribuição da nossa amostra.