## Conceitos de Probabilidade e Distribuições

Nesta aula você vai aprender:

* Os principais conceitos de Probabilidade para a Análise Exploratória de Dados
* O que são e como empregar distribuições de Probabilidades
* Teoremas fundamentais de amostras de dados

# Introdução

Na aula anterior você aprendeu a visualizar distribuições dos dados empregando histogramas e gráficos de densidade. Você deve ter notado também a importância da distribuição para o entendimento dos dados e análise gráfica permite distinguir entre distribuições simétricas, em calda à direita (assimetria negativa, com a média é em geral menor que a mediana) ou à direita etc. Algumas dessas distribuições são bastante comuns e aparecem com frequência em uma série de problemas práticos, como a *distribuição normal* que você já deve ter ouvido falar em algum momento no seu dia-a-dia, na escola ou no trabalho. A Análise Exploratória dos Dados busca encontrar certos padrões nos dados e entender sobre essas distribuições ajudará a você a encontrar essa 'estrutura' dos dados.

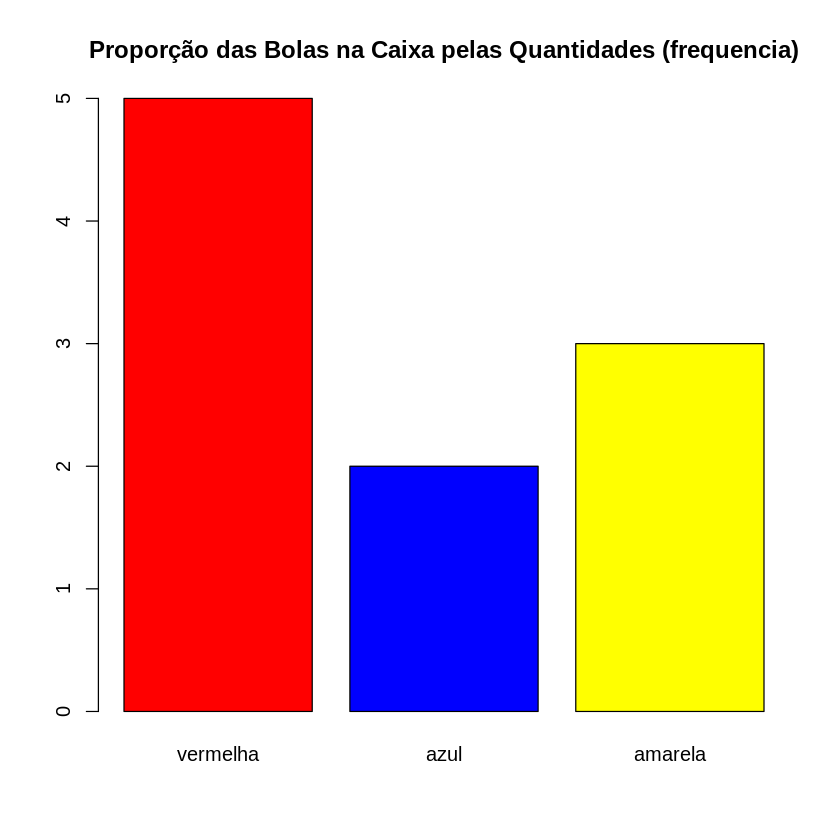
Um caso particular quando analisamos dados é que muitas vezes estaremos lindando com amostras e, como veremos, isso é muito mais comum do que parece. É natural então nos perguntarmos *quanto dos padrões e da estrutura que observada em uma amostra refletem de fato um padrão e a estrutura dos dados?* e veremos aqui resultados que garantem essa *inferência* a partir do que você observa em amostras.

# Probabilidade

Probabilidade é toda uma área que envolve uma série de problemas, alguns bastante difíceis envolvendo análise combinatória e outras áreas da matemática. Certamente você já ouviu ou mesmo aplicou intuitivamente cálculos de probabilidades em jogos de azar, como calcular a chance de ganhar na loteria ou obter um *black-jack* em um jogo do poker, o que é uma aplicação comum de probabilidades. Nosso interesse aqui, entretanto, é mais reduzido e vamos nos concentrar nos conceitos de probabilidades que mais interessam para a Análise Exploratória de Dados e, dentre eles,, encontra-se o conceito das Distribuições de Probabilidades.

Imagine uma caixa com 10 bolas de bilhar, 5 vermelhas, 2 azuis e 3 amarelas. Abaixo você pode ver representada a frequencia, isto é quantidade, de cada uma das cores na caixa.

barplot(c(5,2,3),  
 names.arg=c('vermelha','azul','amarela'),  
 col=c('red','blue','yellow'))  
title('Proporção das Bolas na Caixa pelas Quantidades (frequencia)')



Imagine agora que você vai tirar uma bola da caixa sem olhar para bola que irá retirar. Qual cor você imagina será mais provável de você retirar da caixa? Se você respondeu *vermelha* você certamente já tem a intuição do que é a probabilidade de um evento. A probabilidade, em uma perspectiva que podemos dizer *frequentista*, considera a frequencia (as quantidades) que ocorrem de cada valor, sendo o caso mais provável a cor que aparece em maior proporção. No caso do nosso exemplo, a probabilidade de você tirar cada uma das cores é dada por:

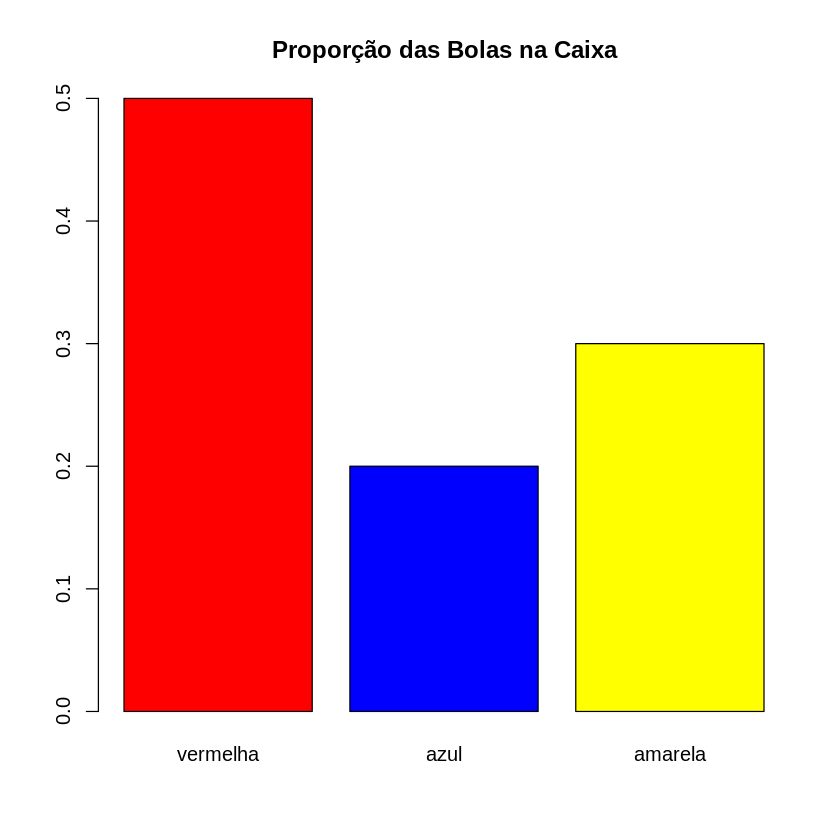
E podemos então dizer que a probabilidade de um evento, um valor , é dada por:

Você pode notar também que a soma da probabilidade de todos os casos é 1.

Este é um resultado não ocorre só para o nosso exemplo. Em qualquer caso a soma das probabilidades de todos eventos possíveis é 1, e podemos escrever isso de modo geral como:

Podemos agora, de forma semelhante ao que fizemos com as quantidades, representar a *proporção* ou a probabilidade de cada cor, dividindo a frequencia de cada cor pelo total de bolas na caixa.

barplot(c(5/10,2/10,3/10),  
 names.arg=c('vermelha','azul','amarela'),  
 col=c('red','blue','yellow'))  
title('Proporção das Bolas na Caixa')



Você acaba de descobrir uma função de probabilidades e isso vai ajudar você a entender as distribuições de probabilidade.

# Distribuições de Probabilidade

As distribuições de probabilidade são um conceito fundamental em estatística. Para dados univariados, funções de distribuição de probabilidades fornecem um modelo de distribuição dos dados, e ela ainda são a base para o cálculo de intervalos de confiança e os testes de hipóteses, desempenhando um papel fundamental na inferência estatística, simulações etc.

A definição matemática de uma **função de probabilidade discreta**, , é uma função com as seguintes propriedades:

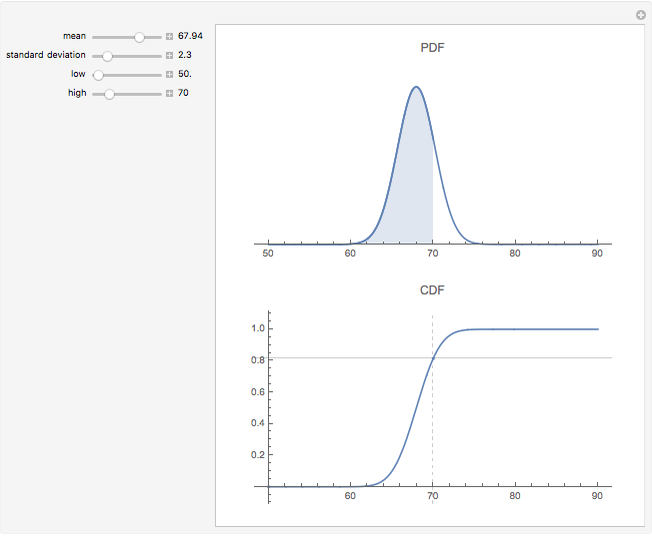
1. Fornece a probabilidade de que possa assumir um determinado valor:
2. é não negativo para todo valor .
3. A soma de para todos os valores possíveis de é 1

E como consequência de 2. e 3. temos para .

Pelo nosso exemplo da caixa com bolas coloridas é exatamente o que você esperava encontrar, não é mesmo?

Uma função de probabilidade discreta é uma função que pode assumir um número discreto de valores (não necessariamente finitos). No nosso exemplo, existem somente 3 casos correspondente às 3 cores na caixa. Mas de modo análogo podemos também **definir funções de probabilidades contínuas**, e com suas correspondentes distribuições.

Para essas funções podemos criar gráficos que são bastante úteis para entendermos os dados envolvidos. Um gráfico que exibe a probabilidade para cada valor de é um **gráfico de densidade de probabilidade** e ainda é comum empregarmos um **gráfico de densidade de probabilidade acumulada** exibindo os valores de para todo .



Brown, R. J. [Connecting the CDF and the PDF](http://demonstrations.wolfram.com/ConnectingTheCDFAndThePDF/) *Wolfram Demonstrations Project*.

Essas funções são normalmente conhecidas como **PDF** (do inglês, Probability Density Function) e **CDF** (do inglês, Cumulative Density Function). E talvez você queira explorar aqui em [Connecting the CDF and the PDF](https://demonstrations.wolfram.com/ConnectingTheCDFAndThePDF/) como essas funções se relacionam.

É importante você notar que, assim como no caso discreto, em que a somatória de todos os valores de probabilidade possíveis é 1, para o caso contínuo,   
, isso significa que a área sob a curva da PDF tem valor 1.

# Função de Probabilidade Discreta: Distribuição Binomial

Considere que você vai jogar uma moeda e deseja saber qual a probabilidade de obter 7 caras ou mais em 20 lançamentos. Esses lançamentos de moeda seguem uma distribuição conhecida como Binomial e podemos empregar a fórmula:

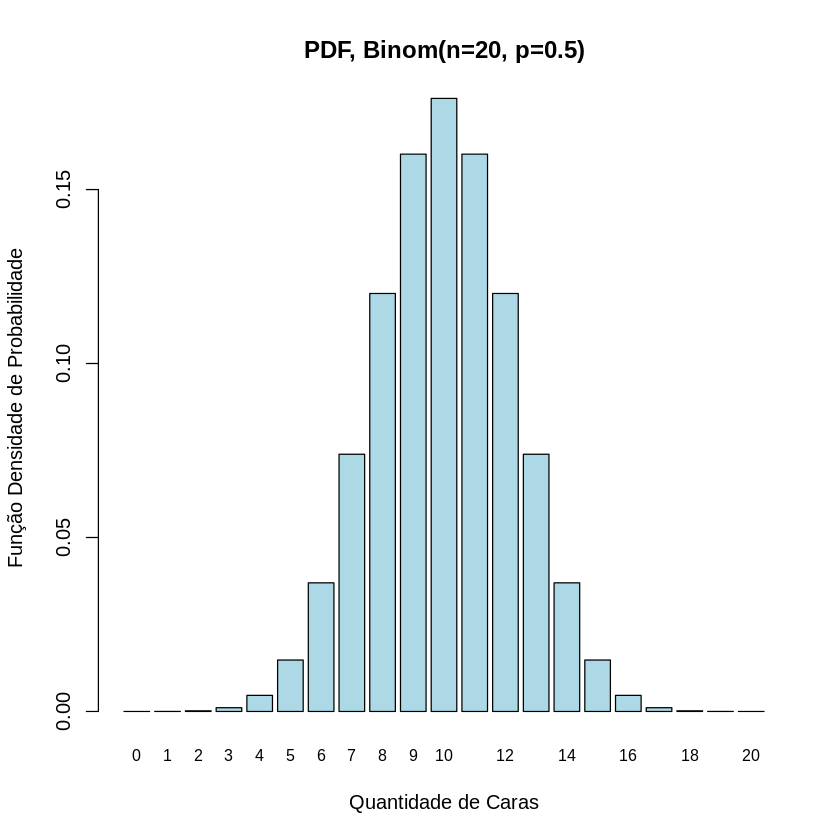
ou resumidamente,

para calcular a probabilidade de até 7 caras, onde é a probabilidade de se obter cara ( se a moeda é não viciada) e o número de lançamentos.

Você não precisa se preocupar compreender completamente essa função agora, mesmo por quê empregaremos o R para fazer esses cálculos para você. O seu objetivo agora é entender *o quê* podemos fazer se conhecemos essa função? Você pode, por hora, pensar nessa função como um fórmula, havendo uma fórmula para cada tipo de distribuição diferente. Entender o que podemos fazer essas funções vai então motivá-lo a compreendê-las melhor.

Podemos então construir os gráficos PDF e CDF dessa distribuição e isso ajudará você a compreender e explorar esse problema.

n = 20  
p = 0.5  
  
prob = c(0)  
  
pmf <- function(k, n, p){  
 return ( factorial(n) / ( ( factorial(k) \* factorial(n - k) ) ) \* p\*\*k \* (1-p)\*\*(n-k) )  
}  
  
for (k in c(0:n)){  
 prob[k+1] = pmf(k, n, p)  
}  
  
barplot(prob,  
 names.arg = seq(0,n,1),  
 xlab = 'Quantidade de Caras',  
 ylab = 'Função Densidade de Probabilidade',  
 col='lightblue',  
 cex.names = 0.8)  
title('PDF, Binom(n=20, p=0.5)')



O gráfico acima exibe a função de densidade de probabilidade, isto é, a probabilidade de obtermos **até** um determinado número de caras ao longo dos 20 lançamentos.

Vamos explorar um pouco esse gráfico. Por exemplo, você pode notar que a probabilidade de termos **até** uma única cara nos lançamentos é muito baixa, afinal você precisaria que os 19 demais lançamentos retornassem todos coroa. A probabilidade aumenta até 10, o quê também é esperado, já que em média, sendo a uma moeda sem vício, a chance de se obter cara é de 50% e 10 corresponderia à metade dos 20 lançamentos.

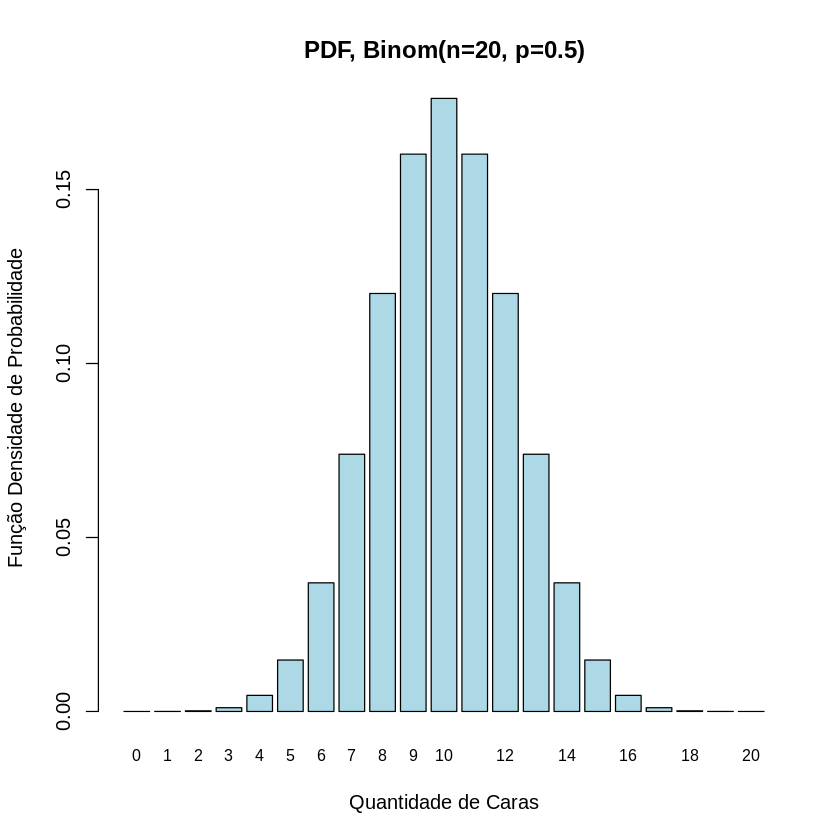
Mas para muitas distribuições, como esta, o R já fornece a função de probabilidades.

dbinom(x, size, prob, log = FALSE)

Veja help('pbinom').

Assim, podemos reescrever o código acima como:

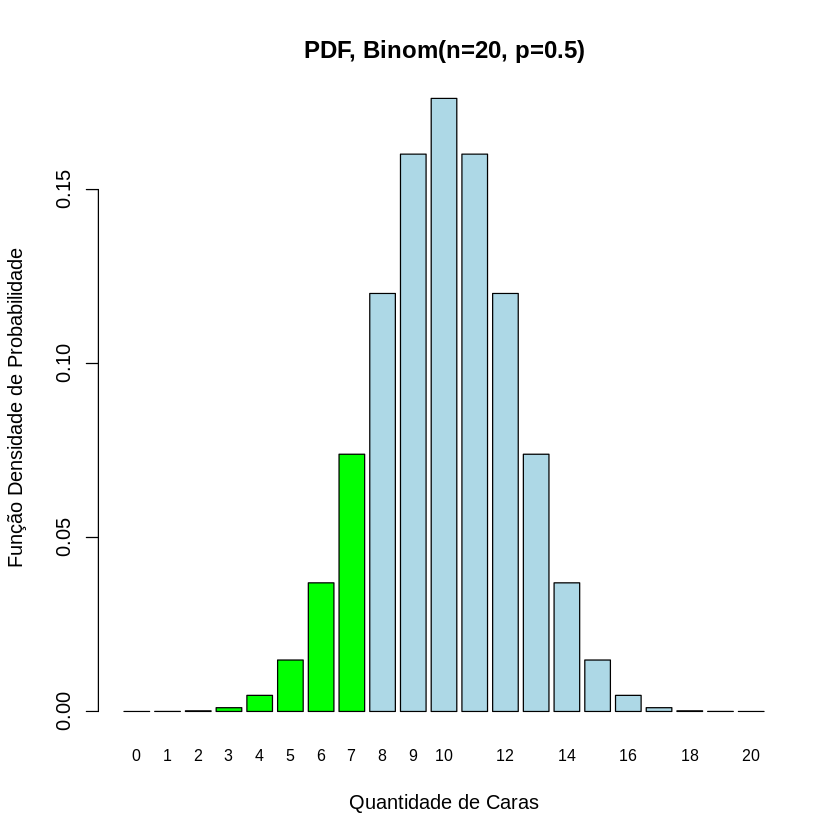
n = 20  
p = 0.5  
  
prob = c(0)  
  
for (k in c(0:n)){  
 prob[k+1] = dbinom(k, n, p)  
}  
  
barplot(prob,  
 names.arg = seq(0,n,1),  
 xlab = 'Quantidade de Caras',  
 ylab = 'Função Densidade de Probabilidade',  
 col='lightblue',  
 cex.names = 0.8)  
title('PDF, Binom(n=20, p=0.5)')



E você verá que outras distribuições a função de **d**istribuição de probabilidades também será **d**, por exemplo dnorm() para a distribuição normal. Mas voltaremos nisso mais adiante.

Mas esse gráfico permite você entender muito mais. A probabilidade de termos até 7 caras nos lançamentos é a soma das probabilidades de termos até 1, 2,... e 7 lançamentos, o que é representado no gráficos pela área ou 'soma' das barras até o valor 7. De fato, você verá mais adiante, que nas distribuições contínuas podemos calcular essa área para obtermos as probabilidades.

n = 20  
p = 0.5  
  
prob = c(0)  
  
for (k in c(0:n)){  
 prob[k+1] = dbinom(k, n, p)  
}  
  
barplot(prob,  
 names.arg = seq(0,n,1),  
 xlab = 'Quantidade de Caras',  
 ylab = 'Função Densidade de Probabilidade',  
 col=c( rep('green',8) , rep('lightblue',n-8) ),  
 cex.names = 0.8)  
title('PDF, Binom(n=20, p=0.5)')



Assim, podemos calcular:

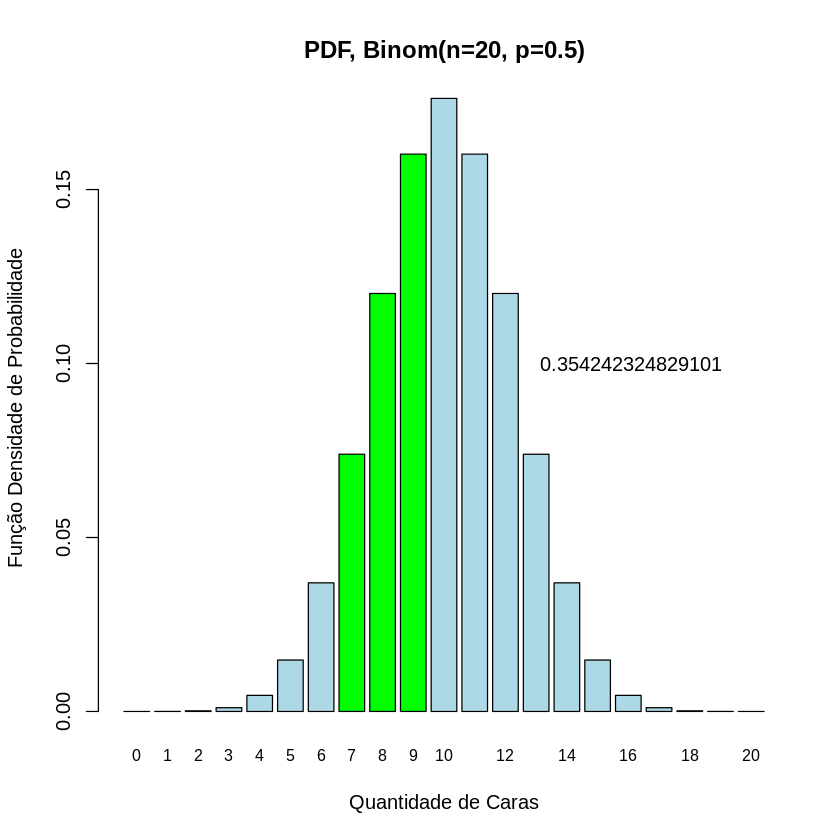
n = 20  
p = 0.5  
  
p7 = 0  
  
for (k in c(0:7)){  
 p7 = p7 + dbinom(k, n, p)  
}  
  
print(p7)

[1] 0.131588

E se você, por exemplo, quiser saber a probabilidade de ter exatamente entre 7 e 9 caras nos lançamentos basta obter a 'soma' (equivalente à área sob a curva em uma distribuição contínua) sob o gráfico entre os valores 7 e 9. E assim, obtemos:

n = 20  
p = 0.5  
  
# Este código, está correto, mas em R podemos substituir por uma única linha  
# pois as operações são aplicadas a todo um vetor  
#----------------------------------------------------------------------------  
# prob = c(0)  
# for (k in c(0:n)){  
# prob[k+1] = dbinom(k, n, p)  
# }  
# Assim, agora que você já entendeu o que estamos fazendo, podemos empregar  
# para obter os valores de dbinom para os valores de c(0:n):  
#----------------------------------------------------------------------------

prob = dbinom(c(0:n), n, p)  
  
# e empregaremos esse modo daqui para diante  
  
p7\_9 = 0  
  
for (k in c(7,8,9)){  
 p7\_9 = p7\_9 + dbinom(k, n, p)  
}  
  
barplot(prob,  
 names.arg = seq(0,n,1),  
 xlab = 'Quantidade de Caras',  
 ylab = 'Função Densidade de Probabilidade',  
 col=c( rep('lightblue',7) , rep('green',10-7), rep('lightblue',n-10) ),  
 cex.names = 0.8)  
title('PDF, Binom(n=20, p=0.5)')  
text(20,0.1,p7\_9)



o gráfico acima empregamos o **gráfico de densidade de probabilidade** com valores de , mas como você notou a soma de várias probabilidades parece ser bastante útil e por isso uma outra forma de representarmos essas probabilidades é empregar um **gráfico de densidade de probabilidade acumulada** exibindo os valores de para todo . Assim o gráfico fornece na posição 7 exatamente o valor de $P(X \le 7) $, isto é, a probabilidade acumulada .

Assim como para funções de distribuição, o R também fornece a função de Função de Probabilidade Acumulada para várias distribuições e, para a distribuição binomial empregamos:

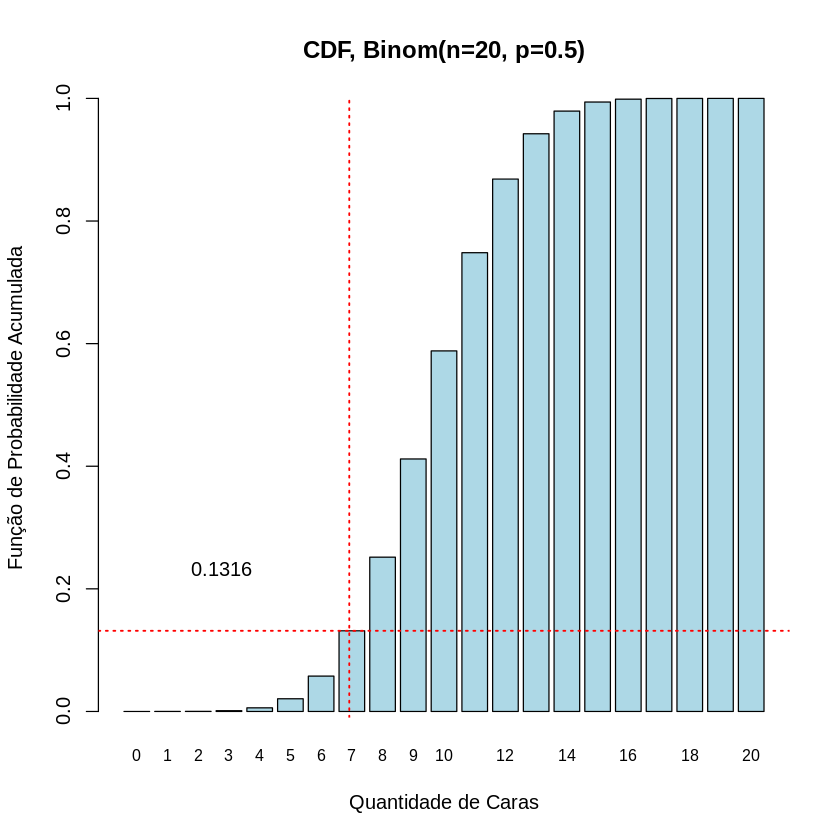
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)

Veja help('pbinom').

Para outras distribuições a função de probabilidade acumulada também será **p**. Por exemplo pnorm() para a distribuição normal.

Assim, podemos obter o gráfico de Probabilidade Acumulada:

n = 20  
p = 0.5  
  
prob = pbinom(c(0:n), n, p)  
  
barplot(prob,  
 names.arg = seq(0,n,1),  
 xlab = 'Quantidade de Caras',  
 ylab = 'Função de Probabilidade Acumulada',  
 col='lightblue',  
 cex.names = 0.8)  
title('CDF, Binom(n=20, p=0.5)')  
  
text(4,pbinom(7, n, p)+0.1,round(pbinom(7, n, p),4))  
abline(h=pbinom(7, n, p),col='red',lty='dotted',lwd=1.5)  
abline(v=7+2,col='red',lty='dotted',lwd=1.5)



A função de probabilidade acumulada permite então obter diretamente a probabilidade de até 7 caras . Assim o cálculo é obtido diretamente:

pbinom(7, n, p)

[1] 0.131588

E ainda podemos obter a probabilidade de termos exatos 7 a 9 caras:

pbinom(9, n, p) - pbinom(6, n, p)

[1] 0.3542423

Embora este seja um exemplo simples, uma série de problemas práticos estão associados a esta mesma função de distribuição de probabilidades. Por exemplo você pode pensar na probabilidade de se obter até 10 peças com defeito em um lote de 100 peças adquiridas sabendo que a média de defeitos do fabricante é de 1% nas peças produzidas. Ou, para estimar atrasos ou o retrabalho em desenvolvimento de software nos próximos meses, obtendo a probabilidade de ser necessário mais de 5 correções de programas de software a cada lote 100 desenvolvidos por uma fábrica de software que apresenta em média 5% de retrabalho com esse fornecedor.

# Função de Probabilidade Contínua: Distribuição Normal

Também bastante comum é a correspondente contínua da distribuição binomial, a distribuição normal. Essa distribuição é bastante popular e talvez você até mesmo já tenha ouvido falar dela. Sua importância e popularidade reside no fato de estar associada a uma série de fenômenos naturais e bastante comuns. Por exemplo medidas como altura, peso, circunferência toráxica ou o número do calçado das pessoas seguem uma distribuição normal (dependendo do contexto, como na física e engenharia, podemos usar o termo Gaussiana). Você poderia, então, empregar isso para definir as melhores quantidades de cada número de sapato ou de tamanhos de camisa a serem fabricados, uma vez que a procura desses produtos depende das medidas das pessoas que vão adquirir esses produtos (e de fato é em geral assim que é feito!). Mas não são somente medidas do ser humano que seguem essa distribuição e podemos também encontrar essa mesma distribuição no peso de sementes na agricultura, notas nos resultados do vestibular ou de um concurso, velocidade dos veículos em uma estrada, medidas do coeficiente de inteligência (*QI*) ou ainda no peso de bebês recém-nascidos.

Diferentemente da distribuição binomial que envolve a probabilidade de valores discretos (inteiros, como as quantidades de lançamentos de uma moeda) a distribuição normal está associada a valores contínuos. O peso e altura de uma pessoa, ou o peso de um recém-nascido, pode assumir quaisquer valores em um intervalo contínuo. Por exemplo, medindo-se o peso de vários bebês recém-nascidos você deve encontrar uma média em torno de 3Kg, mas pode encontrar *qualquer* valor em torno dessa média, por exemplo no intervalo , e não é difícil imaginar que valores cada vez maiores (ou menores) sejam cada vez menos prováveis.

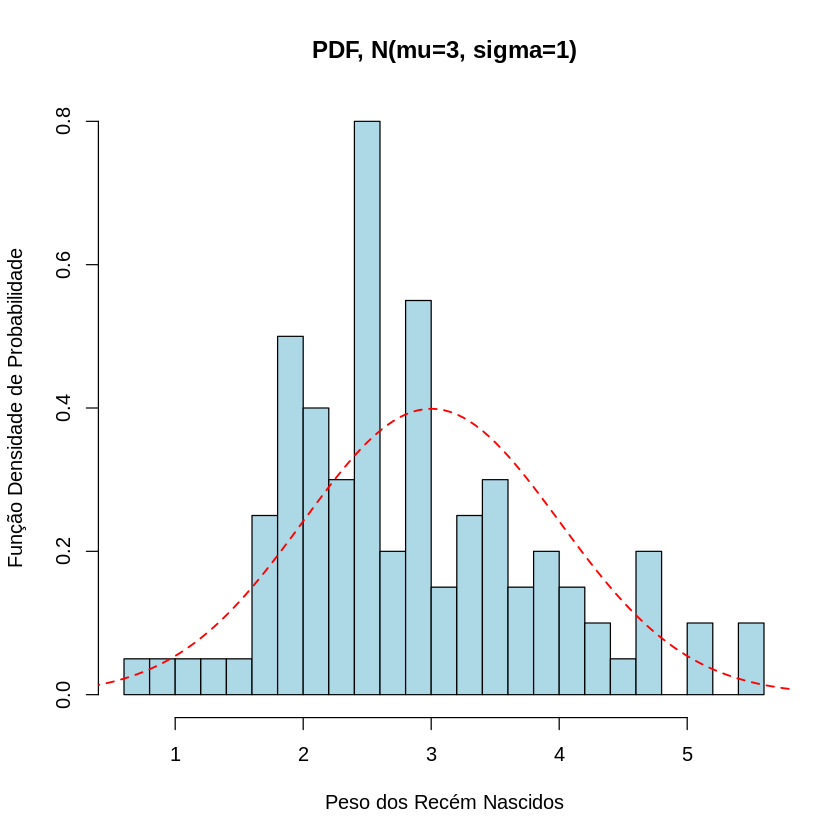
onde é a média da população e o desvio padrão.

Da mesma forma que antes, não é importante que você compreenda essa fórmula agora e você aprenderá como empregar o R para fazer esses cálculos. Mas vamos no concentrar novamente em entender *o que* podemos fazer se conhecemos essa função.

Do mesmo modo que na distribuição contínua podemos os gráficos PDF e CDF dessa distribuição.

Sabendo que a média de peso de recém-nascidos é de 3kg com um desvio padrão de 1Kg, empregamos isso para gerar uma amostra hipotética de peso de 100 bebês com uma distribuição normal e comparar isso à distribuição normal ideal.

set.seed(seed=1234) # fixa a semente de geração aleatória  
  
mu = 3  
sigma = 1  
  
valores = rnorm(n=100,mean=mu,sd=sigma)  
  
hist(valores, breaks=20, freq=FALSE, col='lightblue',  
 main = 'PDF, N(mu=3, sigma=1)',  
 xlab = 'Peso dos Recém Nascidos',  
 ylab = 'Função Densidade de Probabilidade')  
  
# add a 'best fit' line  
y = ((1 / (sqrt(2 \* pi) \* sigma)) \*  
 exp(-0.5 \* (1 / sigma \* (seq(0,6,0.05) - mu))\*\*2))  
lines(seq(0,6,0.05), y, col='red',lty='dashed',lwd=1.5)



A função,

rnorm(n, mean = 0, sd = 1)

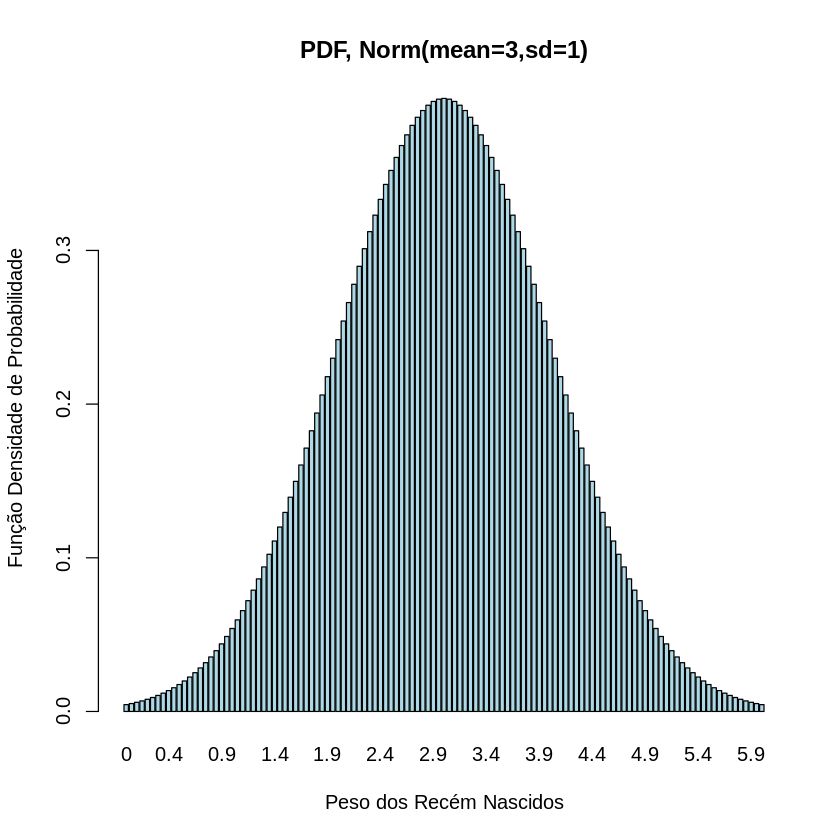
Veja o help(rnorm).

rnorm gera valores aleatórios com a distribuição normal com a média e desvio padrão indicados. Segue o mesmo esquema das funções anteriores, valores **r**andômicos são gerados com rnorm com distribuição normal, **r** para outras distribuições. Por exemplo, rbinom() para a distribuição binomial, mas obviamente você notou que as funções têm parâmetros que diferem para cada tipo de distribuição.

De forma análoga ao caso discreto, a área sob o gráfico da PDF pode ser empregada para o cálculo da probabilidade de um evento e você pode obter a probabilidade de nascimento de um bebê de até 2kg obtendo a área sob a curva até o valor 2.

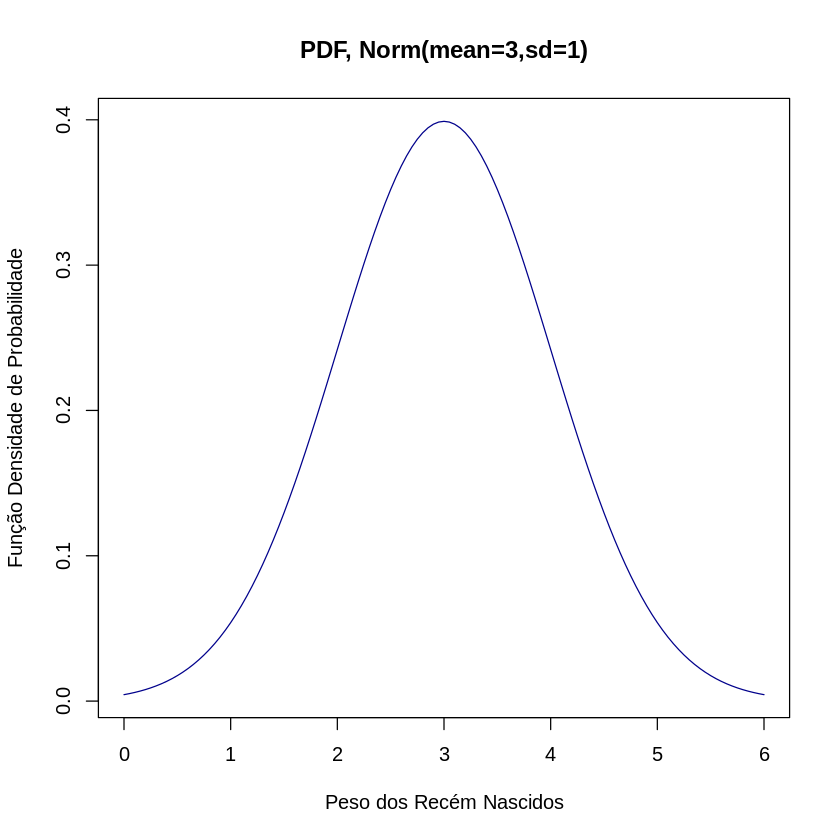
No gráfico abaixo empregamos a função de densidade de probabilidade (PDF) da distribuição normal dnorm() de modo análogo ao que empregamos anteriormente para a distribuição binomial, com dbinom(). É importante que você note que cada tipo de distribuição apresenta parâmetros diferentes. Enquanto, no caso discreto, a distribuição binomial tinha parâmetros como a quantidade de eventos e sua probabilidade, a distribuição normal tem como parâmetros a média e o desvio padrão.

prob = c(0)  
  
mu = 3  
sigma = 1  
  
prob = dnorm(seq(0,6,0.05), mean=mu, sd=sigma)  
  
barplot(prob,  
 names.arg = seq(0,6,0.05),  
 xlab = 'Peso dos Recém Nascidos',  
 ylab = 'Função Densidade de Probabilidade',  
 col='lightblue')  
title('PDF, Norm(mean=3,sd=1)')



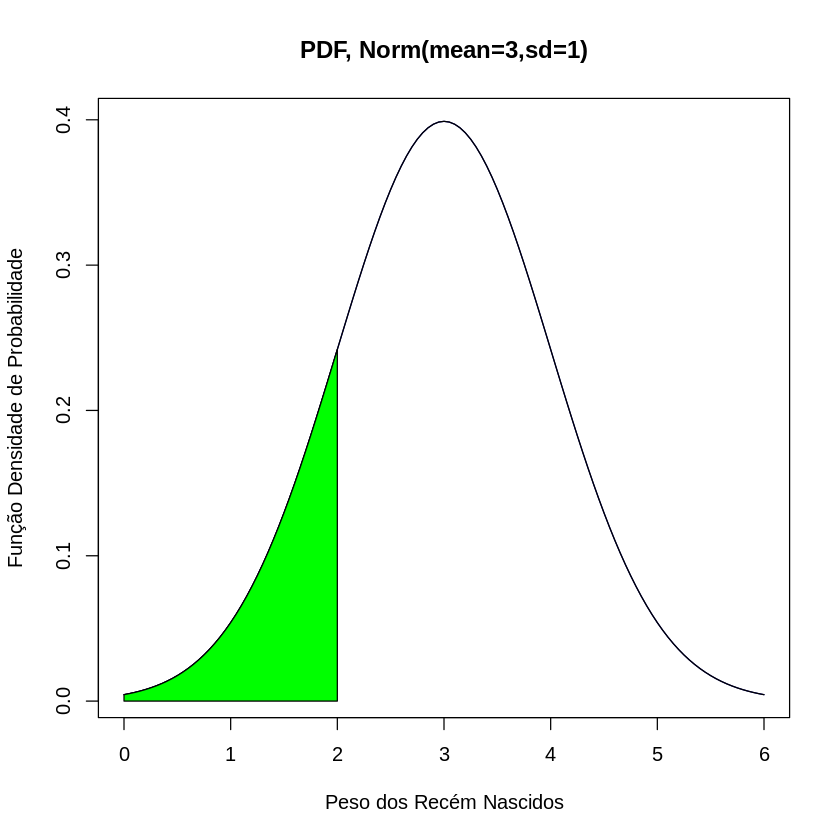
Mas sendo valores contínuos é mais adequado neste caso empregarmos a função plot() no lugar de um histograma.

prob = c(0)  
  
mu = 3  
sigma = 1  
  
prob = dnorm(seq(0,6,0.05), mean=mu, sd=sigma)  
  
plot(seq(0,6,0.05),  
 prob,  
 type='l',  
 xlab = 'Peso dos Recém Nascidos',  
 ylab = 'Função Densidade de Probabilidade',  
 col='darkblue')  
title('PDF, Norm(mean=3,sd=1)')



Agora, de forma análoga ao caso discreto, a área sob o gráfico da PDF pode ser empregada para o cálculo da probabilidade de um evento e você pode obter a probabilidade de nascimento de um bebê de até 2kg obtendo a área sob a curva até o valor 2.

prob = c(0)  
  
mu = 3  
sigma = 1  
  
x =seq(0,6,0.05)  
prob = dnorm(x, mean=mu, sd=sigma)  
  
plot(x,  
 prob,  
 type='l',  
 xlab = 'Peso dos Recém Nascidos',  
 ylab = 'Função Densidade de Probabilidade',  
 col='darkblue')  
title('PDF, Norm(mean=3,sd=1)')  
  
lb = 0 # limite inferior  
ub = 2 # limite superior  
  
i <- x >= lb & x <= ub  
lines(x, prob) # desenha as linhas limite  
polygon(c(lb,x[i],ub), c(0,prob[i],0), col="green") # preenche a área sob a curva



O cálculo dessa área não é tão simples quanto a soma dos valores das probabilidades discretas que fizemos anteriormente. Mas você pode lançar mão função de probabilidades acumulada. Assim como no caso discreto a função **p**:

pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

fornece o valor da função de probabilidade acumulada e podemos então calcular,

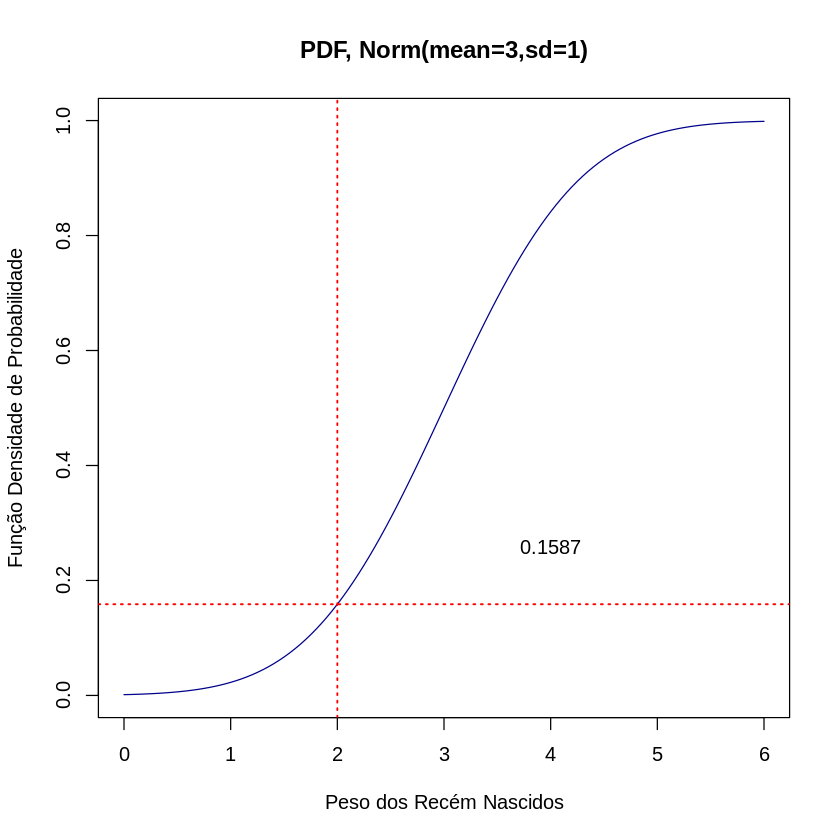
para uma distribuição normal com média 3 e desvio padrão 1:

pnorm(2,mean=3,sd=1)

[1] 0.1586553

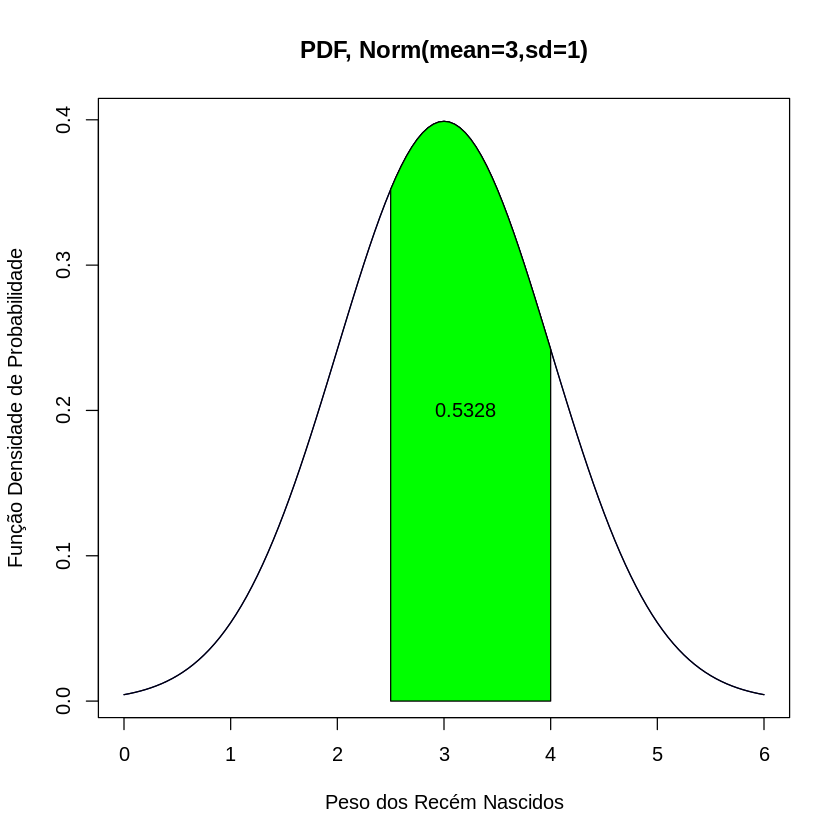
E também construir o gráfico de CDF correspondente:

prob = c(0)  
  
mu = 3  
sigma = 1  
  
x =seq(0,6,0.05)  
prob = pnorm(x, mean=mu, sd=sigma)  
  
plot(x,  
 prob,  
 type='l',  
 xlab = 'Peso dos Recém Nascidos',  
 ylab = 'Função Densidade de Probabilidade',  
 col='darkblue')  
title('PDF, Norm(mean=3,sd=1)')  
  
text(4,pnorm(2,mean=3,sd=1)+0.1,round(pnorm(2,mean=3,sd=1),4))  
  
abline(h=pnorm(2,mean=3,sd=1),col='red',lty='dotted',lwd=1.5)  
abline(v=2,col='red',lty='dotted',lwd=1.5)



Como no caso discreto podemos também calcular a probabilidade de intervalos, como a probabilidade de que o recém-nascido tenha exatamente entre 2.5 e 4.0kg.

prob = c(0)  
  
mu = 3  
sigma = 1  
  
x =seq(0,6,0.05)  
prob = dnorm(x, mean=mu, sd=sigma)  
  
plot(x,  
 prob,  
 type='l',  
 xlab = 'Peso dos Recém Nascidos',  
 ylab = 'Função Densidade de Probabilidade',  
 col='darkblue')  
title('PDF, Norm(mean=3,sd=1)')  
  
lb = 2.5 # limite inferior  
ub = 4 # limite superior  
  
i <- x >= lb & x <= ub  
lines(x, prob) # desenha as linhas limite  
polygon(c(lb,x[i],ub), c(0,prob[i],0), col="green") # preenche a área sob a curva  
  
p\_intervalo = pnorm(4,mean=3,sd=1) - pnorm(2.5,mean=3,sd=1)  
text(3.2,0.2, round(p\_intervalo,4))



pnorm(4,mean=3,sd=1) - pnorm(2.5,mean=3,sd=1)

[1] 0.5328072

# PDF, CDF e ICDF

Há ainda uma outra função importante. A ICDF é a função de distribuição acumulada inversa e dá o valor associado a uma certa probabilidade acumulada. Pegue o exemplo anterior, com 0.15 de chance (probabilidade) até que peso podemos esperar dos bebês?

qnorm(0.15,mean=3,sd=1)

[1] 1.963567

Não é surpresa que o valor é bastante próximo do peso de 2Kg já que essa é a probabilidade de crianças até 2Kg que obtivemos foi:

pnorm(2,mean=3,sd=1)

[1] 0.1586553

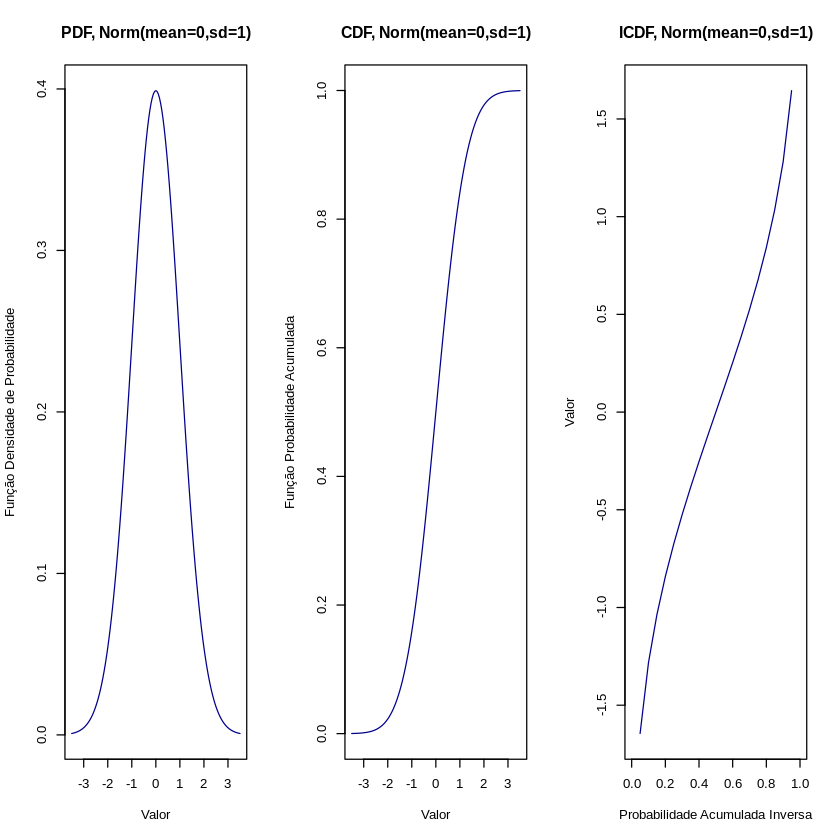
E podemos ainda fazer para obter o valor exato:

qnorm( pnorm(2,mean=3,sd=1) ,mean=3,sd=1)

[1] 2

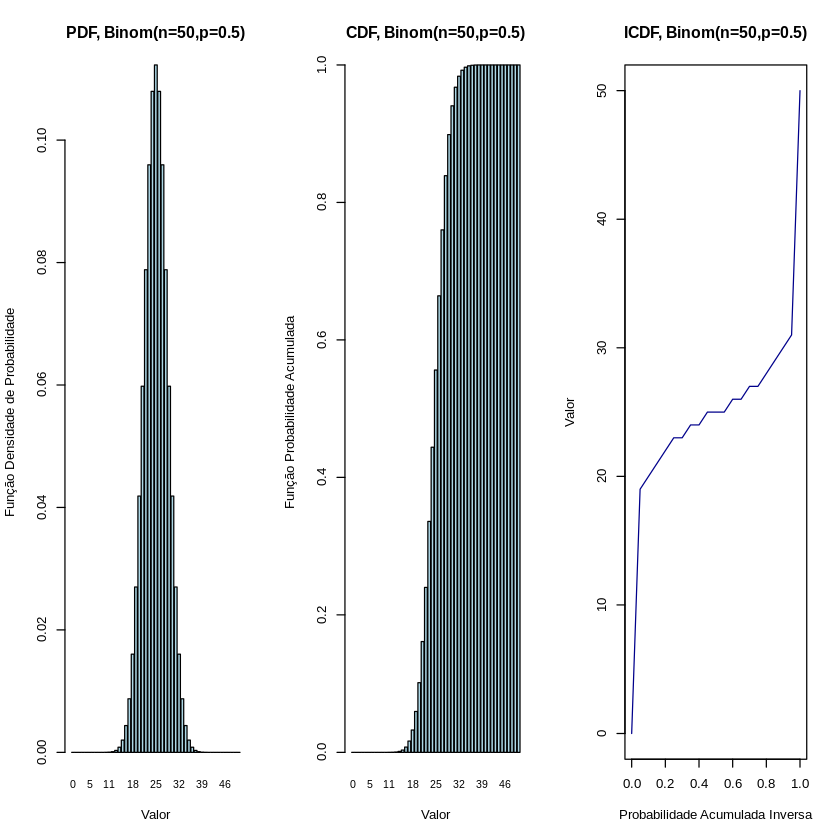
Podemos também construir um gráfico da ICDF e resumir as funções que empregamos até aqui nos seguintes programas:

prob = c(0)  
  
mu = 0 # Aqui empregaremos média 0  
sigma = 1  
  
x = seq(mu - 3.5\*sigma,mu + 3.5\*sigma,0.05)  
  
par(mfrow = c(1, 3))  
  
prob = dnorm(x, mean=mu, sd=sigma)  
plot(x,  
 prob,  
 type='l',  
 xlab = 'Valor',  
 ylab = 'Função Densidade de Probabilidade',  
 col='darkblue')  
title('PDF, Norm(mean=0,sd=1)')  
  
prob = pnorm(x, mean=mu, sd=sigma)  
plot(x,  
 prob,  
 type='l',  
 xlab = 'Valor',  
 ylab = 'Função Probabilidade Acumulada',  
 col='darkblue')  
title('CDF, Norm(mean=0,sd=1)')  
  
prob = seq(0,1,0.05)  
valor = qnorm(prob, mean=mu, sd=sigma)  
plot(prob,  
 valor,  
 type='l',  
 ylab = 'Valor',  
 xlab = 'Probabilidade Acumulada Inversa',  
 col='darkblue')  
title('ICDF, Norm(mean=0,sd=1)')



Ou para o caso discreto,

n = 50 # Aqui com 50 lançamentos  
p = 0.5  
  
prob = c(0)  
  
par(mfrow = c(1, 3))  
  
x = c(0:n)  
prob = dbinom(x, n, p)  
  
barplot(prob,  
 names.arg = x,  
 xlab = 'Valor',  
 ylab = 'Função Densidade de Probabilidade',  
 col='lightblue',  
 cex.names = 0.8)  
title('PDF, Binom(n=50,p=0.5)')  
  
x = c(0:n)  
prob = pbinom(x, n, p)  
  
barplot(prob,  
 names.arg = x,  
 xlab = 'Valor',  
 ylab = 'Função Probabilidade Acumulada',  
 col='lightblue',  
 cex.names = 0.8)  
title('CDF, Binom(n=50,p=0.5)')  
  
prob = seq(0,1,0.05)  
valor = qbinom(prob, n, p)  
  
plot(prob,  
 valor,  
 type='l',  
 ylab = 'Valor',  
 xlab = 'Probabilidade Acumulada Inversa',  
 col='darkblue')  
title('ICDF, Binom(n=50,p=0.5)')



# Resumo das Funções de Probabilidade em R

Como vimos você encontra funções equivalentes em R para:

* Funções (de Densidade) de Probabilidade
* Funções de Probabilidade Acumulada
* Geração aleatória de valores

para as distribuições binomial e normal:

dbinom(x, size, prob, log = FALSE)  
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)  
rbinom(n, size, prob)  
  
dnorm(x, mean = 0, sd = 1, log = FALSE)  
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)  
rnorm(n, mean = 0, sd = 1)

Mas existem outros tipos de distribuições e ainda uma outra função, inversa da função de probabilidade acumulada (retorna o valor de uma variável para uma determinada probabilidade) e que veremos mais adiante. O quadro abaixo resume essas funções e as principais distribuições.

E talvez você prefira empregar a seguinte dica prática:

**Dica prática:**

"d" = **d**ensidade de probabilidade "p" = **p**robabilidade acumulada "q" = **q**uartis (probabilidade acumulada inversa) "r" = **r**andom values

# Exemplo: Preços da Gasolina

Uma empresa de pagamentos eletrônicos divulgou um levantamento do preço da gasolina no país. O preço médio encontrado foi de R$ 4.651 com um desvio padrão de 0.198. Os dados levaram em conta cerca de 20 mil postos de combustível. De acordo com eles, o Acre tem a gasolina mais cara (R$ 5.115 / l), e Santa Catarina, a mais barata (média de média de R$ 4.185 / l). Considerando uma distribuição normal dos preços da gasolina podemos então fazer uma série de análises.

* Qual a probabilidade de encontrarmos o preço da gasolina abaixo de R$ 5.000 o litro?

mu = 4.651  
sigma = 0.198  
  
pnorm(5, mean=4.651, sd=sigma)

[1] 0.9610182

* Qual a probabilidade de encontrarmos o preço da gasolina maior que o preço médio do Acre, de R$ 5.115 o litro?

mu = 4.651  
sigma = 0.198  
  
1 - pnorm(5.115, mean=4.651, sd=sigma)

[1] 0.009553562

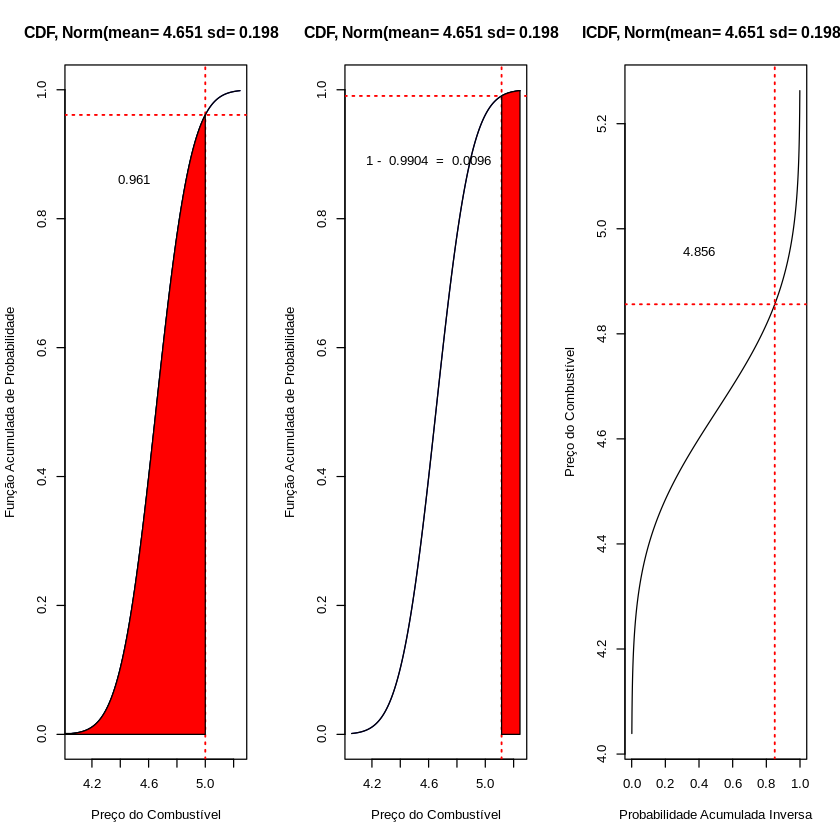
* Qual o preço máximo você colocaria no litro da gasolina para garantir estar entre os 85% preços mais baixos e assim garantir uma maior procura do seu produto?

mu = 4.651  
sigma = 0.198  
  
qnorm(0.85, mean=4.651, sd=sigma)

[1] 4.856214

Graficamente podemos representar as soluções como abaixo. Mas veja que demos as respostas sem a necessidade de exibi-las ou solucionar graficamente. De fato, não é necessário e é bastante complicado produzir gráficos desse modo, e esses gráficos estão aqui apenas para ajudá-lo a compreender a solução. Note também que empregamos do CDF (o gráfico de probabilidade acumulada) e a região colorida da curva está indicando apenas o lado a que corresponde a probabilidade indicada e não a área da curva, o que só empregamos para gráficos de densidade de probabilidade (PDF).

prob = c(0)  
  
mu = 4.651  
sigma = 0.198  
  
x =seq(mu - 3\*sigma,mu + 3\*sigma,0.001)  
prob = pnorm(x, mean=mu, sd=sigma)  
  
par(mfrow = c(1, 3))  
  
plot(x,  
 prob,  
 type='l',  
 xlab = 'Preço do Combustível',  
 ylab = 'Função Acumulada de Probabilidade',  
 col='darkblue')  
t = paste('CDF, Norm(mean=', mu ,'sd=', sigma, ')')  
title(t)  
  
text(4.5,pnorm(5, mean=mu, sd=sigma)-0.1,round(pnorm(5, mean=mu, sd=sigma),4))  
  
abline(h=pnorm(5, mean=mu, sd=sigma),col='red',lty='dotted',lwd=1.5)  
abline(v=5,col='red',lty='dotted',lwd=1.5)  
  
lb = 0 # limite inferior  
ub = 5 # limite superior  
  
i <- x >= lb & x <= ub  
lines(x, prob) # desenha as linhas limite  
polygon(c(lb,x[i],ub), c(0,prob[i],0),  
 col="red") # preenche a área sob a curva  
  
plot(x,  
 prob,  
 type='l',  
 xlab = 'Preço do Combustível',  
 ylab = 'Função Acumulada de Probabilidade',  
 col='darkblue')  
t = paste('CDF, Norm(mean=', mu ,'sd=', sigma, ')')  
title(t)  
  
text(4.6,pnorm(5.115, mean=mu, sd=sigma)-0.1,  
 paste('1 - ', round(pnorm(5.115, mean=mu, sd=sigma),4),' = ',  
 round(1-pnorm(5.115, mean=mu, sd=sigma),4)))  
  
abline(h=pnorm(5.115, mean=mu, sd=sigma),col='red',lty='dotted',lwd=1.5)  
abline(v=5.115,col='red',lty='dotted',lwd=1.5)  
  
lb = 5.115 # limite inferior  
ub = max(x) # limite superior  
  
i <- x >= lb & x <= ub  
lines(x, prob) # desenha as linhas limite  
polygon(c(lb,x[i],ub), c(0,prob[i],0),  
 col="red") # preenche a área sob a curva  
  
  
prob = seq(0,1,0.001)  
valor = qnorm(prob, mean=mu, sd=sigma)  
  
plot(prob,  
 valor,  
 type='l',  
 ylab = 'Preço do Combustível',  
 xlab = 'Probabilidade Acumulada Inversa',  
 col='black')  
t = paste('ICDF, Norm(mean=', mu ,'sd=', sigma, ')')  
title(t)  
  
text(0.4,qnorm(0.85, mean=mu, sd=sigma)+0.1,round(qnorm(0.85, mean=mu, sd=sigma),3))  
  
abline(h=qnorm(0.85, mean=mu, sd=sigma),col='red',lty='dotted',lwd=1.5)  
abline(v=0.85,col='red',lty='dotted',lwd=1.5)



# Regra Prática da Distribuição Normal

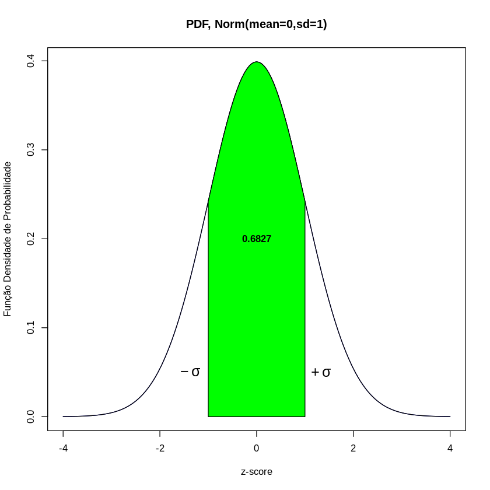
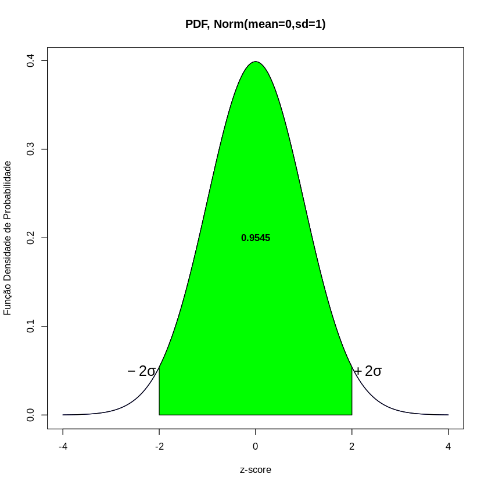
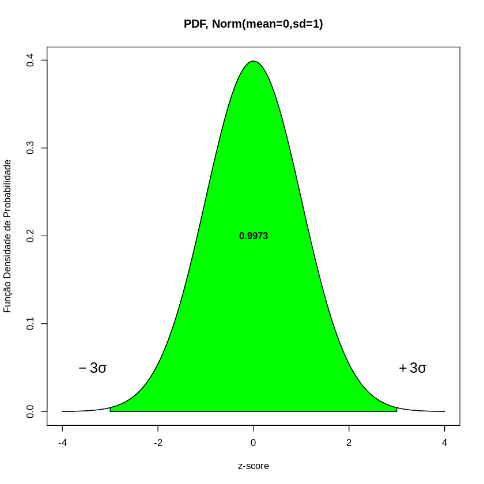
Uma característica importante da distribuição normal (e outras distribuições simétricas) é uma regra empírica que descrevemos a seguir.

De com essa regra, em uma variável aleatória normalmente distribuída, cerca de 68% das observações estarão contidas no intervalo de -1 a +1 desvio-padrão (), 95% das observações estará entre e e cerca de 99.7% das observações estará entre e +3 desvios-padrão.

install.packages('latex2exp')  
library(latex2exp)

Installing package into ‘/usr/local/lib/R/site-library’  
(as ‘lib’ is unspecified)

library(latex2exp)  
  
# par(mfrow = c(1, 3))  
layout.matrix <- matrix(c(1, 1, 1, 0), nrow = 2, ncol = 2)  
  
layout(mat = layout.matrix,  
 heights = c(2, 2), # Heights of the two rows  
 widths = c(2, 2)) # Widths of the two columns  
  
prob = c(0)  
  
mu = 0  
sigma = 1  
  
x =seq(-4,4,0.05)  
prob = dnorm(x, mean=mu, sd=sigma)  
  
plot(x,  
 prob,  
 type='l',  
 xlab = 'z-score',  
 ylab = 'Função Densidade de Probabilidade',  
 col='darkblue')  
title('PDF, Norm(mean=0,sd=1)')  
  
lb = -1 # limite inferior  
ub = 1 # limite superior  
  
i <- x >= lb & x <= ub  
lines(x, prob) # desenha as linhas limite  
polygon(c(lb,x[i],ub), c(0,prob[i],0), col="green") # preenche a área sob a curva  
  
p\_intervalo = pnorm(1,mean=0,sd=1) - pnorm(-1,mean=0,sd=1)  
text(0,0.2, round(p\_intervalo,4), font=2)  
text(-1.4,0.05, TeX('$- \\sigma$'), font=2, cex=1.5)  
text(1.3,0.05, TeX('$+ \\sigma$'), font=2, cex=1.5)  
  
prob = c(0)  
  
mu = 0  
sigma = 1  
  
x =seq(-4,4,0.05)  
prob = dnorm(x, mean=mu, sd=sigma)  
  
plot(x,  
 prob,  
 type='l',  
 xlab = 'z-score',  
 ylab = 'Função Densidade de Probabilidade',  
 col='darkblue')  
title('PDF, Norm(mean=0,sd=1)')  
  
lb = -2 # limite inferior  
ub = 2 # limite superior  
  
i <- x >= lb & x <= ub  
lines(x, prob) # desenha as linhas limite  
polygon(c(lb,x[i],ub), c(0,prob[i],0), col="green") # preenche a área sob a curva  
  
p\_intervalo = pnorm(2,mean=0,sd=1) - pnorm(-2,mean=0,sd=1)  
text(0,0.2, round(p\_intervalo,4), font=2)  
text(-2.4,0.05, TeX('$- 2\\sigma$'), font=2, cex=1.5)  
text(2.3,0.05, TeX('$+ 2\\sigma$'), font=2, cex=1.5)  
  
prob = c(0)  
  
mu = 0  
sigma = 1  
  
x =seq(-4,4,0.05)  
prob = dnorm(x, mean=mu, sd=sigma)  
  
plot(x,  
 prob,  
 type='l',  
 xlab = 'z-score',  
 ylab = 'Função Densidade de Probabilidade',  
 col='darkblue')  
title('PDF, Norm(mean=0,sd=1)')  
  
lb = -3 # limite inferior  
ub = 3 # limite superior  
  
i <- x >= lb & x <= ub  
lines(x, prob) # desenha as linhas limite  
polygon(c(lb,x[i],ub), c(0,prob[i],0), col="green") # preenche a área sob a curva  
  
p\_intervalo = pnorm(3,mean=0,sd=1) - pnorm(-3,mean=0,sd=1)  
text(0,0.2, round(p\_intervalo,4), font=2)  
text(-3.4,0.05, TeX('$- 3\\sigma$'), font=2, cex=1.5)  
text(3.3,0.05, TeX('$+ 3\\sigma$'), font=2, cex=1.5)



Importante você notar que com outras distribuições o resultado vai ser diferente!

# z-score ou escore padrão

O z-score é uma medida importante em estatística para compararmos valores de diferentes conjuntos de dados. Ele nos dá **quantos desvios padrões um dado valor está distante em relação à média**, um z-score positivo indica um valor acima da média (e negativos, abaixo). Ele pode ser calculado como:

onde é a média dos valores e o desvio padrão.

Veja um exemplo. Suponha dois jogadores de poker que apresentam o seguinte histórico de desempenho.

Jogador 1: vence 70% das vezes com um desvio padrão de 20%. Jogador 2: vence 40% das vezes com um desvio padrão de 10%.

Em um determinado campeonato o jogador 1 teve 75% de acertos e o jogador 2, 65%. Qual dos jogadores apresentou nesse campeonato um desempenho acima do esperado (isto é superior ao seu histórico)?

Embora o jogador 1 tenha tido um desempenho melhor, o jogador 2 acertou 25% a mais do que seria esperado já que seu desempenho histórico é de 40% 10%, tendo 2.5 desvios padrão acima da sua média histórica, enquanto o jogador 2 teve um desempenho apenas 0.25 desvios padrão acima da sua média e não foi, portanto, dos seus campeonatos mais felizes ao contrário do jogador 2.

Um **erro comum** é pensarmos que o z-score é uma medida de dispersão de uma amostra. Mas basta voltarmos para definição para vermos que não é, ele é uma medida de **quantos desvios padrões um dado valor está distante em relação à média**. Mas essa medida tem um papel importante nas definições que seguem, da distribuição chi-quadrado até a importante distribuição de t-student.

# Distribuição t-Student

A distribuição t de Student, ou simplesmente t-Student, é uma das distribuições mais utilizadas na estatística com aplicações desde a modelagem estatística até testes de hipóteses que você em outras disciplinas ao longo do curso. De forma bastante simplicada podemos dizer que empregamos distribuições de t-Student quando queremos analisar dados que possuem uma distribuição normal a partir de uma amostra reduzida dos dados. O tamanho da amostra, , é empregado para definirmos um grau de liberdade, , utilizado nos cálculos da distribuição t-Student.

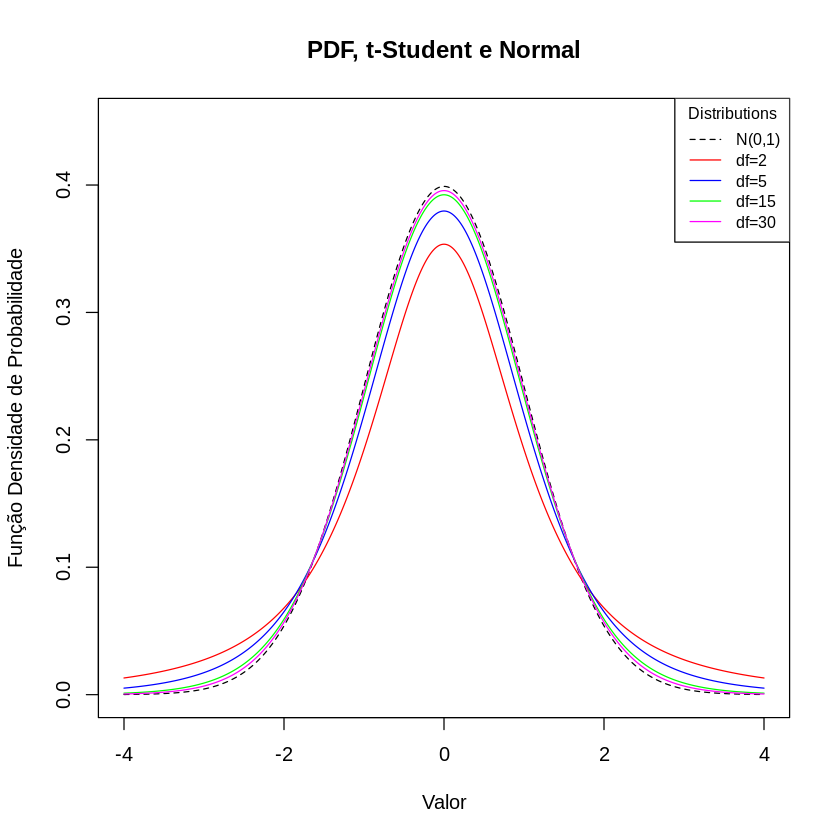
***Grau de liberdade*** *é um conceito bastante complexo de estatística. Considere um conjunto de dados qualquer, os graus de liberdade é o número de valores deste conjunto de dados que podem variar após terem sido impostas certas restrições a todos os valores. Para dar um exemplo, considere que 10 estudantes obtiveram em um teste média 7.0. Assim, a soma das 10 notas deve ser 70 (essa será a restrição). Você terá, então, 10 − 1 = 9 graus de liberdade, pois as nove primeiras notas podem ser escolhidas aleatoriamente, contudo a 10a nota deve ser igual a [70 − (soma das 9 primeiras)]. É algo bem abstrato e aqui, em geral, bastará você pensar em que o grau de liberdade como o número de amostras menos 1. O grau de liberdade aparece nas funções de distribuição em R como df (degrees of freedom).*

A distribuição de t-Student tem uma série de propriedades importantes que você deve saber:

1. Cada grau de liberdade dá origem a uma distribuição t diferente.
2. A função densidade tem a mesma forma em sino da distribuição Normal, mas reflete uma maior variabilidade (com curvas mais alargadas) que é de se esperar em amostras pequenas.
3. A distribuição t-Student se aproxima da normal quando aumenta o número de graus de liberdade.

Na prática a distribuição t-Student é empregada fazermos inferências de uma variável com uma distribuição supostamente normal, mas que temos acesso somente a uma amostra dos dados (e não aos dados de toda a população). Por isso a distribuição t-Student se aproxima da normal a medida que aumentamos o número de graus de liberdade, o que é o mesmo que dizer que aumentamos o número de amostras (lembre: que os graus de liberdade são dados por para amostras!). E você pode observar que amostras já são amostras suficientemente grandes que aproximam bastante bem a distribuição normal.

library(latex2exp)  
  
prob = c(0)  
  
x = seq(-4,4,0.01)  
  
prob = dnorm(x)  
plot(x,  
 prob,  
 type='l',lty=2, lwd=1,  
 xlab = 'Valor',  
 ylab = 'Função Densidade de Probabilidade',  
 col='black',  
 ylim=c(0,0.45))  
title('PDF, t-Student e Normal')  
  
cores = c('darkblue', 'red', 'blue', 'green', 'magenta')  
color = 2  
  
for (df in c(2,5,15,30)){  
prob = dt(x, df=df)  
  
lines(x, prob, col=cores[color])  
color = color + 1  
  
  
}  
  
legend("topright", legend=c('N(0,1)','df=2','df=5','df=15','df=30'),  
 title='Distributions',  
 col=c('black', 'red', 'blue', 'green', 'magenta'), lty = c(2,1,1,1,1), cex=0.8)



# Outras Distribuições

Existem muitas outras distribuições importantes, como você pode ver pelo Resumo das Funções de Probabilidade acima. Mas com o que aprendeu aqui você já terá um ferramental bastante grande e útil para prosseguir e poderá explorar outras distribuições quando precisar.

A ideia central da inferência estatística é que podemos estimar ou *inferir* características de uma população a partir de uma amostra dessa população. Isto é, não precisamos acessar ou medir toda a população! A partir de uma amostra, podemos estimar os parâmetros de uma população, como por exemplo a média e a variância. A Lei dos Grandes números afirma que a medida que crescemos a amostra esses parâmetros da amostra tendem aos mesmos parâmetros da população e Teorema do Limite Central destaca a importância da distribuição normal para a inferência de parâmetros.

# Lei dos Grandes Números e o Teorema Central do Limite

O que podemos dizer sobre a média, a variância ou a distribuição dos dados quando só temos acesso a uma amostra dos dados? Quando pensamos em amostras frequentemente associamos as amostras às pesquisas estatísticas. Pesquisadores selecionam uma amostra de brasileiros (algumas centenas dentre milhões) para uma pesquisa de intenção de voto a fim de inferir um possível resultado nas eleições. Ou ainda, um nutricionista pode selecionar amostras dos alimentos produzidos por uma fábrica para verificar sua qualidade e buscar inferir se a amostra aponta algum problema ou não para os produtos em geral. Mas você pode pensar em amostras de uma forma mais ampla em que, simplesmente, não temos acesso a todos os dados. Por exemplo, sobre pode ter *todas* (e não só uma amostra!) as informações sobre vendas do último ano ou dados de todos os seus clientes, mas o que podemos dizer sobre as vendas do próximo mês ou de novos clientes? Obter respostas como essas é objetivo da *inferência estatística* (que não cobriremos aqui) e em boa parte são possíveis graças a dois resultados da teoria de probabilidades que você verá a seguir. A Lei dos Grandes números afirma que a medida que crescemos a amostra esses parâmetros da amostra tendem aos mesmos parâmetros da população e Teorema do Limite Central destaca a importância da distribuição normal para a inferência de parâmetros. Esses resultados asseguram a possibilidade de empregarmos dados de amostras para estimarmos parâmetros como médias e distribuições da população dos dados.

## Lei dos Grandes Números

Uma das principais atividades da estatística é a generalização dos resultados obtidos em uma amostra para população de onde obtivemos a amostra. Isto é, você faz uma inferência dos resultados da população com base nos dados da amostra. Por exemplo, você pode inferir a média de peso dos recém-nascidos com base em uma amostra de, digamos, 30 bebês, e as distribuições de probabilidade têm um papel fundamental nessa tarefa de inferência.

A Lei dos Grandes Números e o Teorema Central do Limite são dois teoremas fundamentais da teoria da probabilidade para inferência estatística. Eles impactam no tamanho amostral e como usamos a distribuição normal para fazer inferências.

A **Lei dos Grandes Números (LGN)**, afirma que a média dos resultados de um experimento realizado repetidas vezes *tende* a se aproximar do valor esperado à medida que novas tentativas se sucedem. Isto, é, a média amostral *tende* ou **converge** à média populacional à medida que mais experimentos são realizados. Em outras palavras, quanto maior a amostra, mais os resultados se aproximam dos resultados da população.

Uma demonstração simples pode ser feita pelo jogo de moedas. Em uma moeda não viciada sabe-se que há uma probabilidade de 0.5 do resultado ser cara (valor 1) e a mesma probabilidade 0.5 de ser coroa (valor 0). Então, se assumirmos como o número de caras de cada experimento :

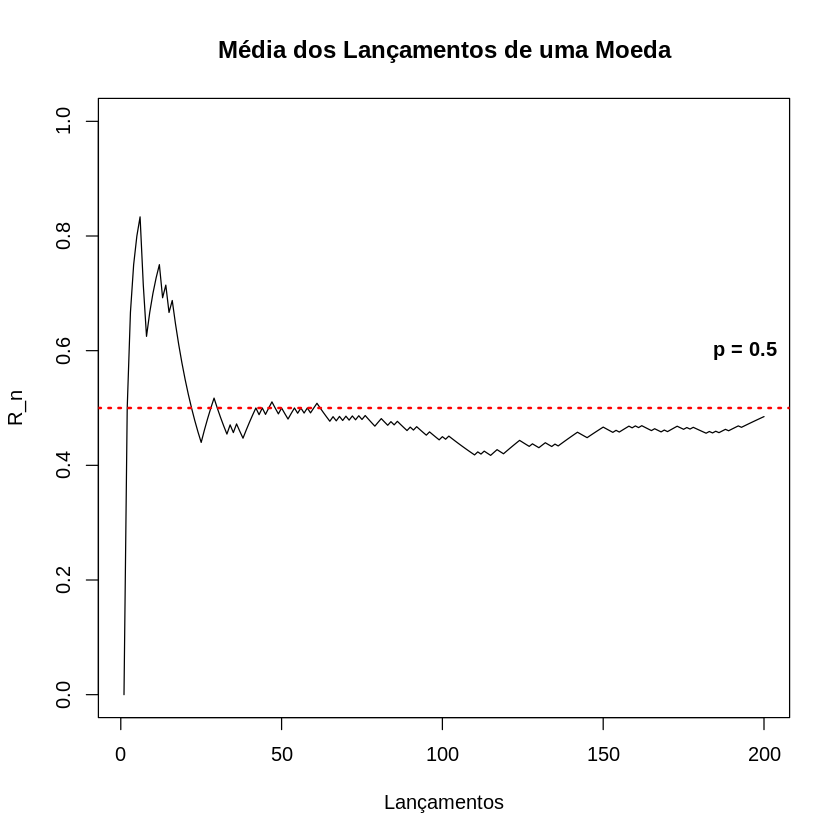
A razão de caras no momento é dada por:

e, pela **Lei dos Grandes Números (LGN)**, quanto maior o número de experimentos mais próximo da probabilidade 0.5 será o nosso resultado:

Você pode observar essa *convergência* para o valor 0.5 no gráfico a seguir onde simulamos 200 lançamentos de uma moeda.

set.seed(1234)  
lanctos = as.numeric(runif(200) >= 0.5)  
print(lanctos[1:100])  
  
media = c(0)  
  
for (i in 1:200){  
 media[i] = sum(lanctos[1:i]) / i  
}  
  
plot(media,type='l',main='Média dos Lançamentos de uma Moeda',  
 ylim=c(0,1),  
 xlab='Lançamentos',  
 ylab='R\_n')  
abline(h=0.5,col='red',lty='dotted',lwd=2)  
text(194,0.6, 'p = 0.5', font=2, cex=1)

[1] 0 1 1 1 1 1 0 0 1 1 1 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1 0 1 0  
 [38] 0 1 1 1 1 0 1 0 1 1 0 0 1 0 0 1 1 0 1 0 1 0 1 1 0 0 0 0 1 0 1 0 1 0 1 0 1  
 [75] 0 1 0 0 0 1 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 0 1



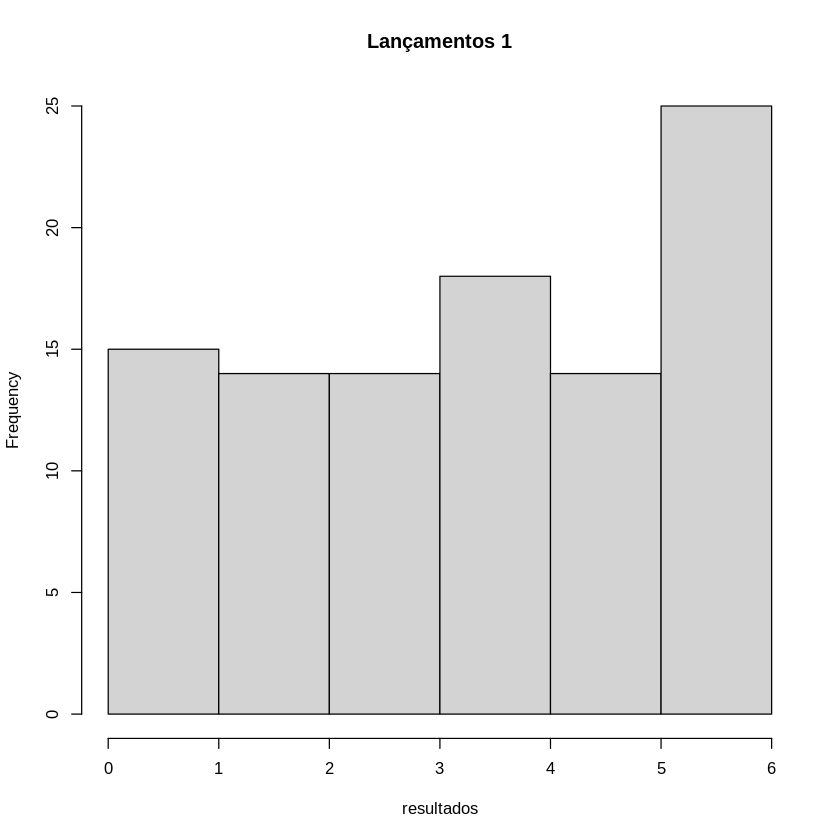
Ah! Você já viu isso não é!? Sim, veja o exercício 2 da aula anterior!

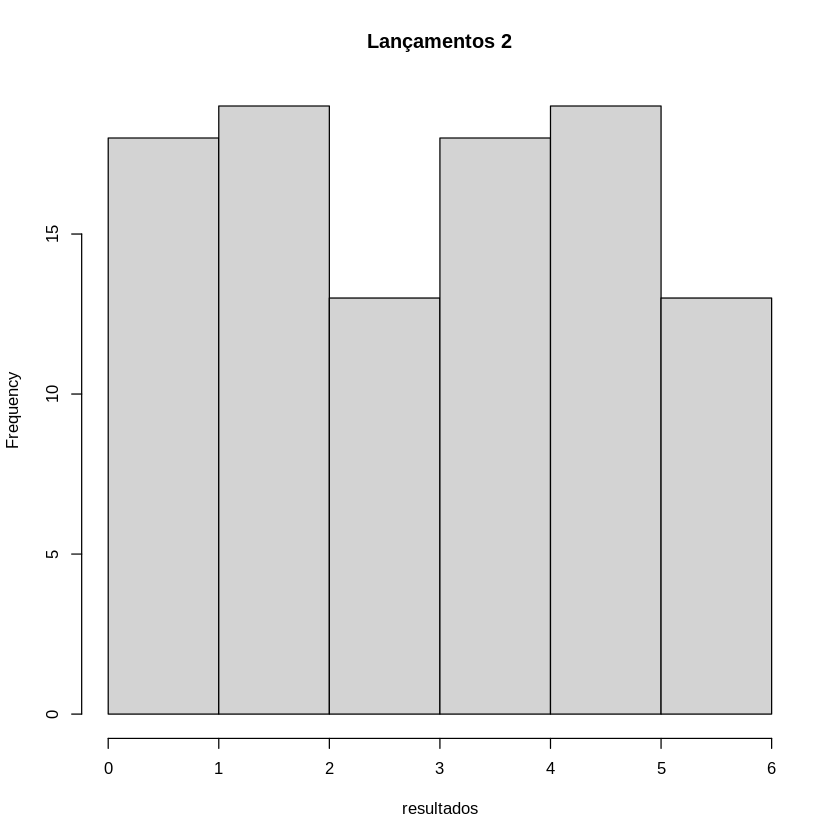
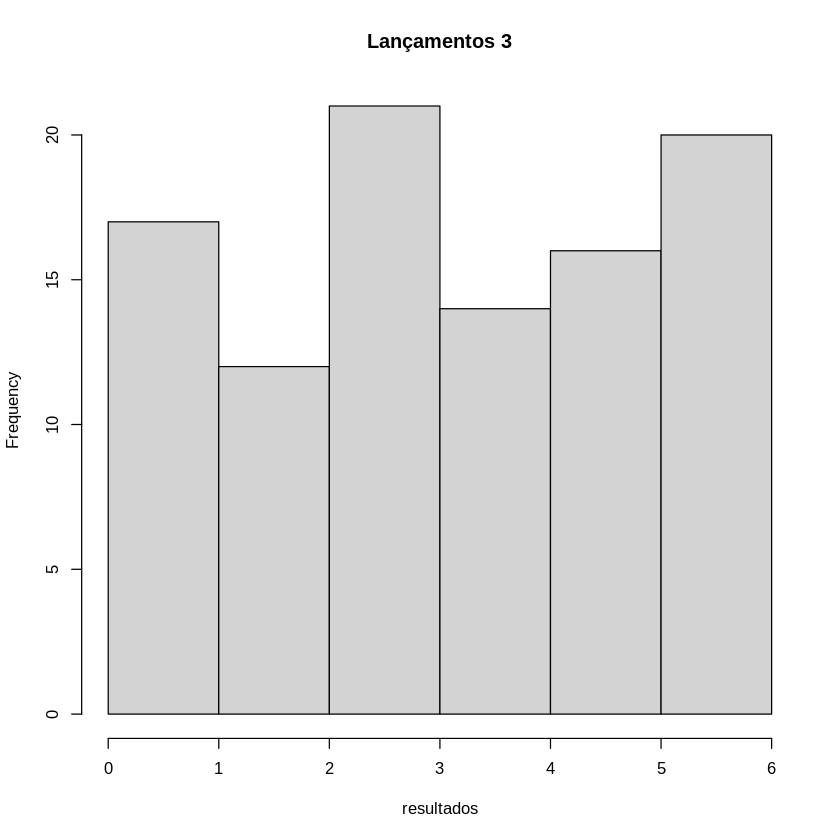
## Teorema Central do Limite

O **Teorema Central do Limite (TCL)** é um outro resultado importante e foi considerado como “Central” pelo matemático George Pólya. Esse teorema afirma que a distribuição da média de um grande número de variáveis aleatórias independentes e identicamente distribuídas será aproximadamente normal, independentemente da distribuição subjacente (dessas variáveis).

Por isso você verá que a distribuição normal é utilizada para inferência e em muitos testes estatísticos.

layout.matrix <- matrix(c(1, 1, 1, 0), nrow = 2, ncol = 2)  
  
layout(mat = layout.matrix,  
 heights = c(2, 2), # Heights of the two rows  
 widths = c(2, 2)) # Widths of the two columns  
  
set.seed(1234)  
hist(sample(1:6,100, replace=TRUE),  
 breaks=c(0,1,2,3,4,5,6),  
 main='Lançamentos 1',  
 xlab='resultados')  
  
set.seed(1984)  
hist(sample(1:6,100, replace=TRUE),  
 breaks=c(0,1,2,3,4,5,6),  
 main='Lançamentos 2',  
 xlab='resultados')  
  
set.seed(7777)  
hist(sample(1:6,100, replace=TRUE),  
 breaks=c(0,1,2,3,4,5,6),  
 main='Lançamentos 3',  
 xlab='resultados')



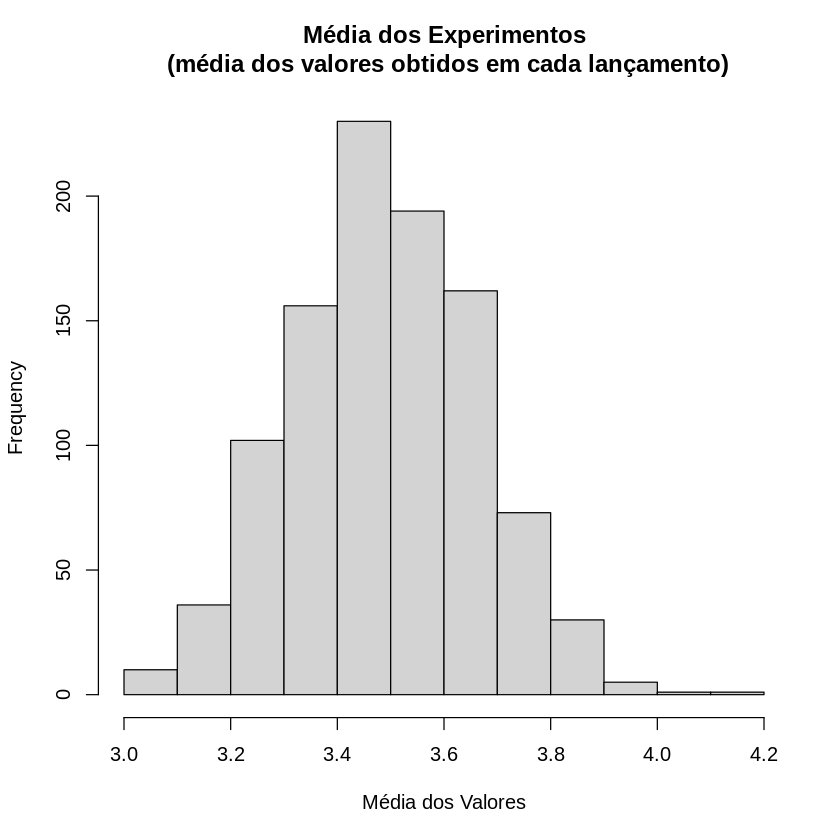


Vamos considerar o experimento acima de lançamento de um dado não viciado. Como cada resultado de 1 a 6 tem probabilidade 1/6 de ocorrer **essa é uma distribuição uniforme** (importante: note aqui, todos os valores tem a mesma probabilidade e, portanto, o gráfico de densidade de probabilidade tem um formato de caixa ou retângulo, e também denominamos essa distribuição de retangular).

Acima você encontra os resultados de 3 experimentos de 100 lançamentos cada. Note o formato das distribuições em 'caixa'.

O que o **Teorema Central do Limite** afirma é que se fizermos vários desses experimentos, a média desses experimentos terá uma distribuição normal. Veja abaixo a simulação de 1000 experimentos de 100 lançamentos cada e a distribuição da média obtida>

set.seed(1234)  
media = c(0)  
  
for (i in 1:1000){  
 media[i] = mean(sample(1:6,100, replace=TRUE))  
}  
  
hist(media,  
 main='Média dos Experimentos\n (média dos valores obtidos em cada lançamento)',  
 xlab='Média dos Valores')



Um resultado surpreendente não acha? Sim, veja que essa **distribuição normal** ocorre independentemente da distribuição subjacente das variáveis que apresentam, neste caso, uma distribuição uniforme, mas poderiam apresentar qualquer outra forma de distribuição.

# Para Saber Mais

1. Explore aqui de forma interativa a conexão entre as curvas de distribuição de probabilidades e de probabilidade acumulada em uma distribuição normal! Brown, R. J.**Connecting the CDF and the PDF** disponível em: <http://demonstrations.wolfram.com/ConnectingTheCDFAndThePDF/> *Wolfram Demonstrations Project*.
2. Que tal acessar nossa biblioteca digital e explorar mais esses conceitos? O livro de **Devore, Jay L.**, **Probabilidade e Estatística para Engenharia e Ciências**, é uma referência no ensino de Probabilidade e Estatística e lá você poderá rever vários pontos apresentados aqui com uma abordagem bastante formal e completa. Leia o Capítulo 3 **Variáveis aleatórias discretas e distribuições de probabilidades** (Introdução e itens 3.1-3.2) e o Capítulo 4 **Variáveis aleatórias contínuas e distribuições de probabilidade** (Introdução e itens 4.1-4.3).

Aqui o Link da biblioteca virtual: <https://integrada.minhabiblioteca.com.br/#/books/9788522128044>

1. Você também pode consultar **Freire, Sergio Miranda. Bioestatística Básica**. disponível em: <http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html>.

Os capítulos 10 e 11 tratam de distribuições de probabilidade e você pode achar esse material mais acessível.

# Referências

1. Devore, Jay L. Probabilidade e Estatística para Engenharia e Ciências. Trad. Solange Aparecida Visconte. Cengage, 2018.
2. Navarro, Danielle, **Learning Statistics with R**, disponível em: <https://learningstatisticswithr.com/> ( LSR version 0.6 (pdf) ). Acesso: 26/02/2021. Alternativamente em formato bookdown: <https://learningstatisticswithr.com/book/> Acesso: 07/03/2021.
3. Wickham, H., Grolemund, G. **R for Data Science**. O'Reilly Media, Inc., 2016.
4. Glen, S. **Welcome to Statistics How To!** From StatisticsHowTo.com: Elementary Statistics for the rest of us! disponível em: <https://www.statisticshowto.com/> Acesso: 09/03/2021.
5. Brown, R. J.**Connecting the CDF and the PDF** disponível em: <http://demonstrations.wolfram.com/ConnectingTheCDFAndThePDF/> *Wolfram Demonstrations Project*. Published: November 2 2007 Acesso: 09/03/2021.
6. Freire, Sergio Miranda. **Bioestatística Básica**. disponível em: <http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html>. Acesso: 10/04/2021.