Formação Cientista de Dados

Dia 02 - Populações e Inferência Estatística

Vítor Wilher

Cientista de Dados | Mestre em Economia

Plano de Voo

Introdução

Variáveis aleatórias discretas

Variáveis aleatórias contínuas

Amostragem de uma população

Famílias de Distribuições

O Teorema Central do Limite

Inferência Estatística

Introdução

Nas seções anteriores do nosso **Curso de Formação Cientista de Dados**, vimos uma introdução à exploração e tratamento de dados. A partir dessa seção, faremos uma iniciação ao processo de inferência estatística propriamente dito. **Inferência estatística**, em termos simplificados, é o processo de formar julgamento sobre uma determinada população com base em uma amostra dessa população. Aqui, nós descrevemos populações e amostras de populações usando a linguagem da probabilidade.[1]

[1] Essa seção procura introduzir o tema de *inferência estatística* de modo despreocupado, baseado sobretudo em Verzani [2014]. Para uma leitura intermediária do assunto, ver Meyer [2011], que tem uma boa introdução à probabilidade. Para uma leitura completa, ver Casella and Berger [2016].

Introdução

De modo a fazer inferência estatística com base em dados nós utilizamos um **modelo probabilístico** para os dados. E aqui devemos diferenciá-lo de um **modelo determínistico**. Modelos determinísticos são aqueles que estipulam que as condições sob as quais um experimento seja executado determinam o resultado do experimento. Por exemplo, se introduzirmos uma bateria em um circuito simples, o modelo matemático que descreveria o fluxo de corrente elétrica seria presumivelmente a *lei de Ohm*, de modo que para obter a intensidade de corrente elétrica, basta que tenhamos a tensão e a resistência elétrica.

Introdução

Considere, por outro lado, um conjunto de dados univariado que consiste em medidas de alguma variável. Um ponto qualquer desse conjunto será a realização de um intervalo de valores. Chamaremos assim a *população* de um variável como sendo a descrição do intervalo de possibilidades para esse valor. Utilizaremos o termo *variável aleatória* para descrever o número aleatório de uma população. Assim, um ponto dentro daquele conjunto de dados nada mais é que a realização de uma determinada variável aleatória.

Introdução

Importante dizer que fazemos uma distinção entre quando nós temos uma variável aleatória observada ou realizada. Uma vez observada, o valor da variável aleatória é conhecido. Antes de ser observada, contudo, ela pode ser qualquer valor dentro daquele intervalo possível da população. Para a maioria dos casos, nem todos os valores da população possuem a mesma *probabilidade* de ocorrerem. Assim, antes de observar uma variável aleatória, nós precisaremos indicar a probabilidade dela exercer um determinado valor ou um intervalo de valores. Nós nos referimos a essa descrição de intervalo e suas respectivas probabilidades como *distribuição de uma variável aleatória*.

Introdução

Por probabilidade, nós queremos dizer algum número entre 0 e 1 que descreve a possibilidade da nossa variável aleatória assumir algum valor. A intuição aqui vem da necessidade de se entender como os números são gerados. Por exemplo, quando jogamos uma moeda para o ar, a probabilidade de dar *cara* ou *coroa* é de 50% para cada. Para situações onde as possibilidades são igualmente prováveis e o total de possibilidades é finito, a definição de probabilidade é dada por

$$P(E) = \frac{\text{Eventos em E}}{\text{Total de Eventos}}. \tag{1}$$

Para essa definição, as seguintes regras serão satisfeitas:

Introdução

- $P(E) \geq 0$ para todos os eventos;
- O evento de todos os possíveis resultados tem probabilidade igual a 1;
- Se A e B são dois eventos disjuntos, então $P(AouB) = P(A) + P(B)$.

Uma consequência matemática disso é que para qualquer dois eventos, $P(AouB) = P(A) + P(B) - P(AeB)$.

Introdução

A tabela a seguir mostra a relação entre gênero e fumo para uma coorte no conjunto de dados `survey` do pacote `MASS`:

```
data(survey, package='MASS')
tbl = xtabs(~Sex + Smoke, data = survey)
tbl
```

```
##          Smoke
## Sex       Heavy Never Occas Regul
##   Female      5    99     9     5
##   Male        6    89    10    12
```

Introdução

Existem 237 participantes no estudo, mas apenas 235 estão representados acima. Se nós selecionarmos um deles *de forma aleatória*, a probabilidade de selecionarmos uma mulher será o número de mulheres dividido pelo número total de pessoas representadas. Isto é,

```
margin.table(tbl, margin=1)
```

```
## Sex
## Female   Male
##    118    117
```

```
sum(tbl[1,])/sum(tbl)
```

```
## [1] 0.5021277
```

Variáveis aleatórias discretas

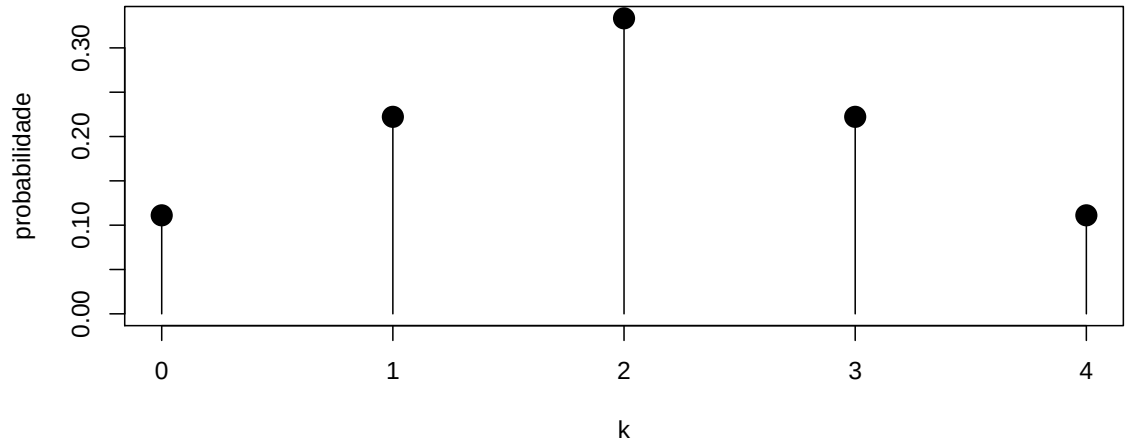
Dados numéricos podem ser discretos ou contínuos, de modo que podemos ter igualmente variáveis aleatórias discretas ou contínuas. Deixemos que X seja uma variável aleatória discreta. Isto é, uma variável aleatória cujos possíveis resultados sejam algo como $\{sim, não\}$ ou $\{0, 1, 2, ...\}$. A extensão de X será o conjunto de todos os k, onde $P(X = k) > 0$. A distribuição de X será uma especificação dessas probabilidades. As regras da probabilidade implicam que distribuições não são arbitrárias, tal que para cada k no intervalo, $P(X = k) > 0$ e $P(X = k) \leq 1$. Ademais, como X tem algum valor, nós teremos $\sum_k P(X = k) = 1$.

Variáveis aleatórias discretas

Especificando uma distribuição

Nós podemos especificar a distribuição de uma variável aleatória discreta primeiro especificando o intervalo de valores e depois associando a cada k um número $p_k = P(X = k)$, de modo que $\sum p_k = 1$ e $p_k \leq 0$.

```
k = 0:4
p = c(1,2,3,2,1); p = p/sum(p)
plot(k, p, type='h', xlab='k', ylab='probabilidade', ylim=c(0,max(p)))
points(k,p,pch=16,cex=2)
```



Variáveis aleatórias discretas

O R possui a função `sample` de modo a gerar observações para uma variável aleatória discreta com uma distribuição específica. Se o vetor k contém os valores amostrados, e p contém as probabilidades dos valores selecionados, assim a função `sample` irá selecionar um dos k valores com a probabilidade especificada por p, como abaixo.

```
k = 0:2
p = c(1,2,1); p = p/sum(p)
sample(k, size=1, prob=p)
```

```
## [1] 2
```

Variáveis aleatórias discretas

Para um determinado conjunto de dados, a média e o desvio-padrão são sumários do centro e da amplitude. Para variáveis aleatórias esses conceitos se transferem, de modo que suas definições são diferentes.

A *média populacional* será denotada por μ. Se X for uma variável aleatória com essa população, a média será também chamada de o *valor esperado de* X, de modo que

$$\mu = E(X) = \sum kP(X = k).$$

Isto é, a média ponderada dos valores no intervalo de X com pesos $p_k = P(X = k)$. Já o *desvio-padrão populacional* será denotado por σ. O desvio-padrão será a raiz quadrada da variância. Se X for uma variável aleatória discreta, sua variância será definida por $\sigma^2 = VAR(X) = E((X - \sigma)^2)$.

Variáveis aleatórias contínuas

Dados contínuos são modelados por variáveis aleatórias contínuas. Devido aos valores possíveis serem contínuos, uma nova forma de definir probabilidades deve ser utilizada. Ao invés de tentar especificar $P(X = k)$ para todo k, defini-se $P(a < X \leq b)$. Isto pode ser feito por meio de uma função, $F(b) = P(X \leq b)$ ou por meio de uma função relacionada $f(x)$ que possui $P(a < X \leq b)$ igualando a área abaixo do gráfico de $f(x)$, entre a e b. Para uma dada variável aleatória X, a função $f(x)$ é referida como a *densidade* de X.

Variáveis aleatórias contínuas

f.d.p. e f.d.a.

Para uma variável aleatória discreta é comum definir a função $f(k)$ por $f(k) = P(X = k)$. De maneira similar, para uma variável aleatória contínua X, é comum definir a densidade de X por $f(x)$. Em ambos os casos, elas serão chamadas de f.d.p. No caso discreto, significa *função de distribuição de probabilidade*, enquanto no caso contínuo significa *função densidade de probabilidade*. A função de distribuição acumulada, por seu turno, será $F(b) = P(X \leq b)$. No caso discreto, será dada por $\sum_{k \leq b} P(X = k)$; no caso contínuo, será a área à esquerda de b sob a densidade $f(x)$.

Variáveis aleatórias contínuas

Os conceitos de média e desvio-padrão também se aplicam a variáveis aleatórias contínuas, embora suas definições requeiram cálculo. A noção intuitiva para a média de X é que esta será o ponto de equilíbrio para a densidade de X. Ficam válidas as mesmas letras. Se X tem, por exemplo, uma distribuição uniforme sobre $[0, 1]$, a média será $\frac{1}{2}$ e o desvio-padrão será aproximadamente 0.289.

Amostragem de uma população

Nosso modelo de probabilidade para um ponto do conjunto de dados será uma observação de uma variável aleatória cuja distribuição é descrita pela população parental. Para performar inferência estatística sobre uma população parental, precisamos de uma *amostra* da população. Isto é, uma sequência de variáveis aleatórias $X_1, X_2, ..., X_n$. Uma sequência será *identicamente distribuída* se cada variável aleatória possuir a mesma distribuição, obviamente. A sequência será independente se ao conhecer o valor de alguma das variáveis aleatórias não implicar em nenhuma informação adicional sobre a distribuição das outras. Uma sequência que é independente e identicamente distribuída (i.i.d.) é chamada de amostra aleatória.

Amostragem de uma população

Para ilustrar, considere jogar uma moeda para o alto n vezes. Se dissermos que X_i é 1 para uma possibilidade e 0, caso contrário, então claramente $X_1, X_2, X_3, ..., X_n$ será uma sequência *i.i.d.*. De modo geral, se nós geramos nossos números aleatórios por meio de uma seleção aleatória a partir de uma população finita, então os valores serão independentes se a amostragem for feita com reposição. Por outro lado, caso não tenhamos reposição, as variáveis aleatórias $X_1, X_2, X_3, ..., X_n$ ainda terão a mesma distribuição, mas serão dependentes.

Amostragem de uma população

A função `sample` tomará amostras de tamanho n de uma distribuição discreta com a especificação `size = n`. Abaixo um exemplo.

```
sample(0:1, size=10, replace=T)
```

```
##  [1] 1 0 0 1 1 1 0 1 1 1
```

Amostragem de uma população

Uma *estatística* é um valor numérico que sumariza uma amostra aleatória. Um exemplo disso é a média amostral $\bar{X} = \sum_{i=1}^{n} \frac{X_n}{n}$. Quando uma estatística depende de uma amostra aleatória, ela também será uma variável aleatória. Daí a descrição $\bar{X}$. A distribuição de uma estatística é chamada de distribuição de amostragem.

Famílias de Distribuições

Em estatística existem distribuições que se apresentam em famílias. Cada família é descrita por meio de uma função que possui um número de parâmetros que caracterizam a distribuição. Por exemplo, a distribuição uniforme é uma distribuição contínua no intervalo $[a, b]$ que atribui probabilidade igual para áreas de tamanho igual no intervalo. Os parâmetros a e b serão os *limites* desse intervalo. A densidade e a população de uma família de distribuições são em geral representadas em termos desses parâmetros.

Famílias de Distribuições

As funções *d*, *p*, *q* e *r*

O R possui quatro tipos de funções para tomar informação sobre uma família de distribuições:

- A função *d* retorna a *f.d.p.* da distribuição;
- A função *p* retorna a *f.d.a.* da distribuição;
- a função *q* retorna os *quantis*;
- a função *r* retorna amostras aleatórias de uma distribuição.

Famílias de Distribuições

Essas funções são usadas de forma similar. Cada família tem um nome e alguns parâmetros. O nome da função é encontrado pela combinação de d, p, q ou r com o nome da família. Os nomes dos parâmetros variam de família para família mas são consistentes com a família.

Por exemplo, a distribuição uniforme sobre $[a, b]$ tem dois parâmetros. O nome da família é `unif`. No R os parâmetros são nomeados como `min` e `max`. Abaixo um exemplo para a distribuição uniforme sobre $[0, 3]$.

Famílias de Distribuições

```r
dunif(x=1, min=0, max=3)
```

```
## [1] 0.3333333
```

```r
punif(q=2, min=0, max=3)
```

```
## [1] 0.6666667
```

```r
qunif(p=1/2, min=0, max=3)
```

```
## [1] 1.5
```

```r
runif(n=1, min=0, max=3)
```

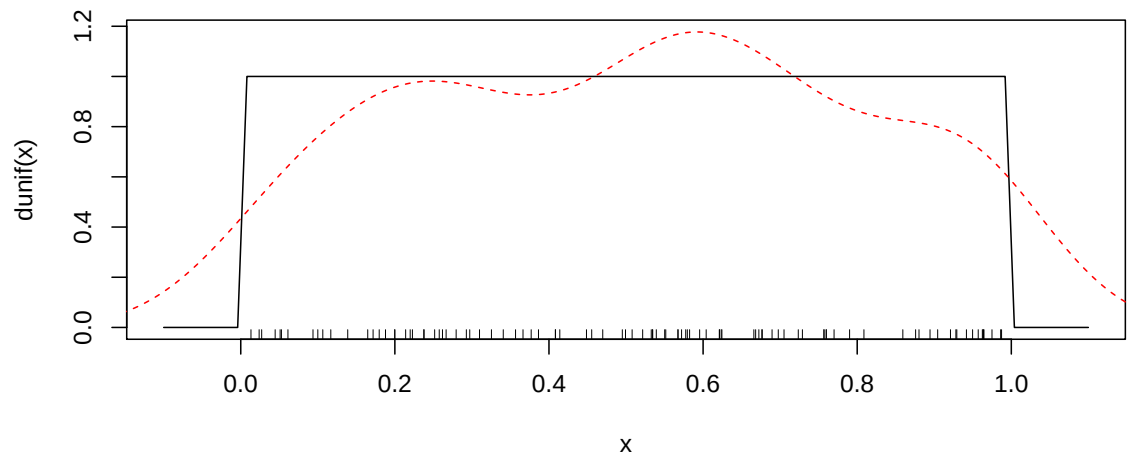
```
## [1] 1.336159
```

Famílias de Distribuições

Relacionando as funções *d* e *r*

Para distribuições contínuas, a função *d* descreve uma densidade teórica. Vamos abaixo relacioná-la com a função *r*.

```
x = runif(100)
d = density(x)
curve(dunif, -0.1,1.1,ylim=c(0,max(d$y,1)))
lines(d, lty=2, col='red')
rug(x)
```



Famílias de Distribuições

Variáveis aleatórias de Bernoulli

Uma variável aleatória de *Bernoulli* X é uma que possui apenas dois valores: 0 e 1. A distribuição de X é caracterizada por $p = P(X = 1)$. Se utiliza `Bernoulli(p)` para se referir a essa distribuição. Em geral, atribui-se *sucesso* quando $X = 1$ e *fracasso* quando $X = 0$. Se jogarmos uma moeda e deixemos que X seja 1 caso dê cara, então X será uma variável aleatória Bernoulli onde o valor de p seria $\frac{1}{2}$ se a moeda for honesta. Uma sequência de moedas jogadas seria uma sequência *i.i.d.* de variáveis aleatórias de Bernoulli, também conhecida como *processo ou ensaio de Bernoulli*. Uma variável aleatória de Bernoulli tem média $\mu = p$ e variância $\sigma^2 = p(1 - p)$. No R, a função `sample` pode ser utilizada para gerar uma amostra aleatória a partir de uma distribuição de Bernoulli. Abaixo um exemplo.

Famílias de Distribuições

```
n = 10; p = 1/4
sample(0:1, size=n, replace=TRUE, prob=c(1-p,p))
```

```
##  [1] 0 1 0 0 0 0 0 0 0 0
```

Famílias de Distribuições

Variáveis aleatórias Binomiais

Uma variável aleatória Binomial X conta o número de sucessos em n processos de Bernoulli. Existem dois parâmetros que descrevem a distribuição de X: o número de processos, n, e a probabilidade de sucesso, p. A distribuição é representada por `Binomial(n,p)`. O intervalo possível para X será $0, 1, ..., n$. A distribuição de X será dada por

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

O termo $\binom{n}{k}$ é conhecido como coeficiente binomial e é definido por

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

Famílias de Distribuições

A notação padrão $n!$ é para o *fatorial* de n ou simplesmente $n * (n-1) * * * 2 * 1$. Por convenção, $0! = 1$. O coeficiente binomial conta o número de maneiras que k objetos podem ser escolhidos de n objetos distintos e é lido como n *escolhe* k. A função `choose` resulta no coeficiente binomial. A média de uma variável aleatóri `Binomial(n,p)` será $\mu = np$ e o desvio-padrão será $\sigma = \sqrt{np(1-p)}$. No R, o nome da família será `binom` e os parâmetros são rotulados como `size = n` e `prob = p`.

Famílias de Distribuições

Joque uma moeda dez vezes para o alto. Deixe que X seja o número de caras. Se a moeda for honesta, X possui uma distribuição `Binomial(10,1/2)`. A probabilidade que $X = 5$ pode ser encontrada diretamente a partir da distribuição com a função `choose`:

```
choose(10,5)*(1/2)^5*(1/2)^(10-5)
```

```
## [1] 0.2460938
```

Famílias de Distribuições

Isso, a propósito, pode ser melhor representado usando a função *d*, isto é, `dbinom`:

```
dbinom(5, size=10, prob=1/2)
```

```
## [1] 0.2460938
```

Famílias de Distribuições

A probabilidade de que seja seis ou menos caras,
$P(X \leq 6) = \sum_{k \leq 6} P(X = k)$, pode ser dada de duas maneiras:

```
sum(dbinom(0:6, size = 10, prob=1/2))
```

```
## [1] 0.828125
```

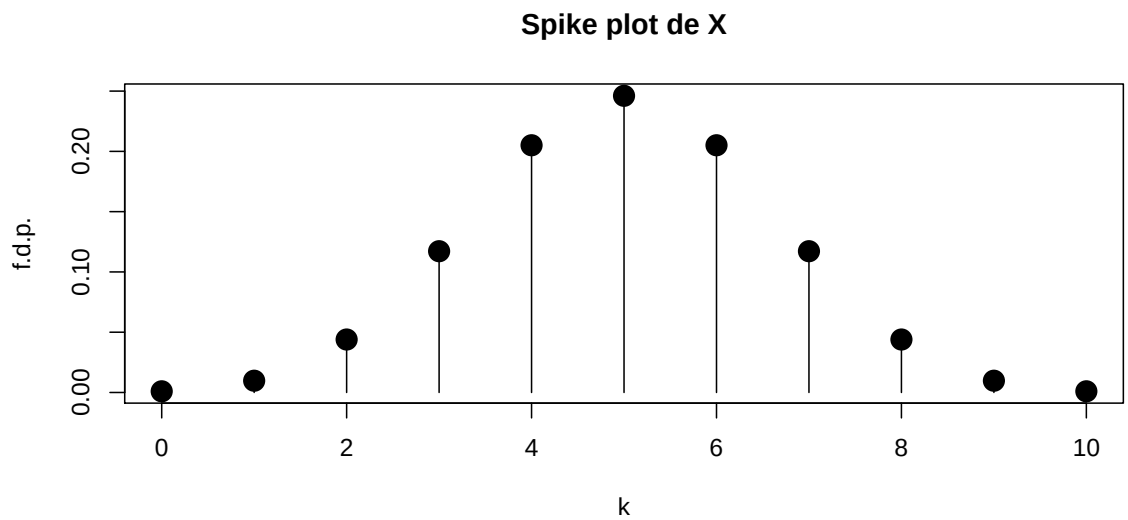
```
pbinom(6, size=10, p=1/2)
```

```
## [1] 0.828125
```

Famílias de Distribuições

Um gráfico da distribuição pode ser gerado usando `dbinom`:

```
n = 10; p = 1/2
heights = dbinom(0:10, size=n, prob=p)
plot(0:10, heights, type='h',
     main='Spike plot de X', xlab='k', ylab='f.d.p.')
points(0:10, heights, pch=16, cex=2)
```



Famílias de Distribuições

Variável aleatória Normal

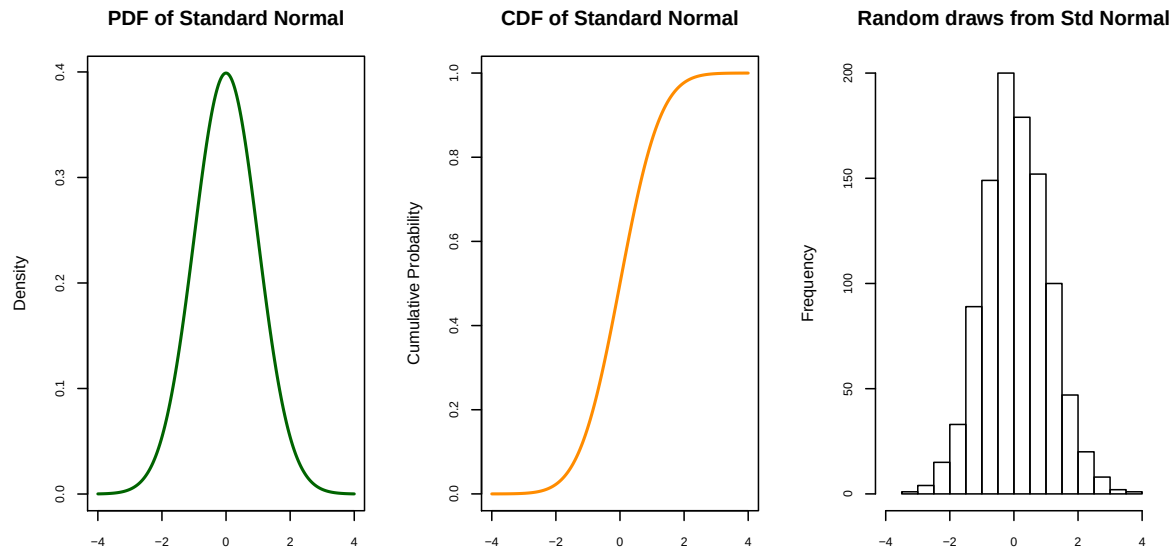
A distribuição normal é uma distribuição contínua que dá sentido à expressão *em forma de sino*. É utilizada para descrever diversas populações na natureza, tal qual a distribuição de alturas, e adicionalmente descreve a distribuição de amostragem de diferentes estatísticas. A distribuição normal é uma família de distribuições com densidade dada por

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\epsilon^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

Os dois parâmetros são a média, μ, e o desvio padrão, σ. Nós utilizamos `Normal`(μ,σ) para representar a distribuição, embora muitos livros utilizem a variância σ^2 para representar o segundo parâmetro. No R, o nome da família é `norm` e os parâmetros são `mean` e `sd`.

Famílias de Distribuições

```
set.seed(3000)
xseq<-seq(-4,4,.01)
densities<-dnorm(xseq, 0,1)
cumulative<-pnorm(xseq, 0, 1)
randomdeviates<-rnorm(1000,0,1)
par(mfrow=c(1,3), mar=c(3,4,4,2))
plot(xseq, densities, col="darkgreen",xlab="", ylab="Density",
     type="l",lwd=2, cex=2, main="PDF of Standard Normal",
     cex.axis=.8)
plot(xseq, cumulative, col="darkorange", xlab="",
     ylab="Cumulative Probability",type="l",lwd=2, cex=2,
     main="CDF of Standard Normal", cex.axis=.8)
hist(randomdeviates, main="Random draws from Std Normal",
     cex.axis=.8, xlim=c(-4,4))
```



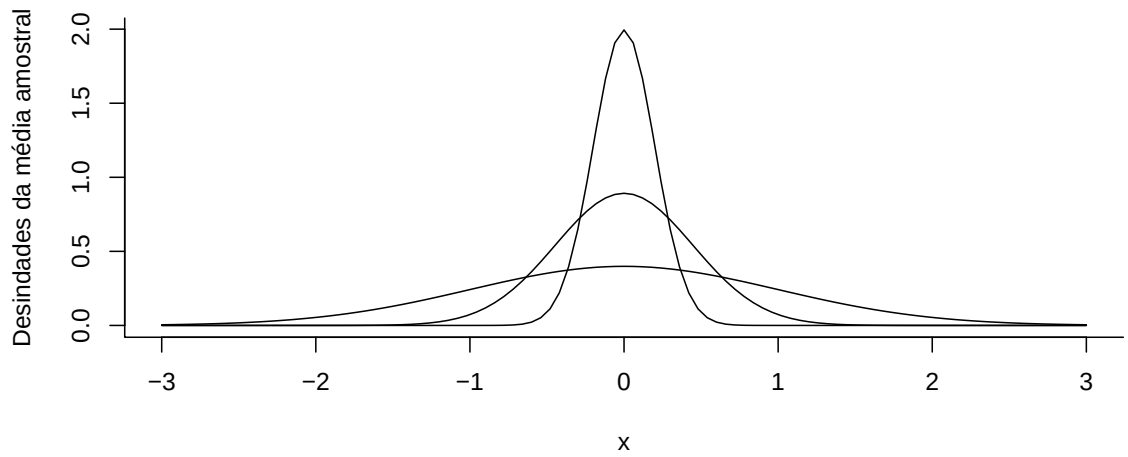
O Teorema Central do Limite

Para uma amostra *i.i.d.* obtida de uma população, a distribuição da média amostral tem valor esperado μ e desvio-padrão $\frac{\sigma}{\sqrt{n}}$, onde μ e σ são parâmetros populacionais. Para um *n grande o suficiente*, a distribuição de amostragem de $\bar{X}$ será normal ou aproxidamente normal.

O Teorema Central do Limite

Quando a amostra $X_1, X_2, ..., X_n$ é desenhada a partir de uma população $N(\mu, \sigma)$, a distribuição de $\bar{X}$ será precisamente a distribuição normal. Abaixo, desenhamos densidades para a população e distribuição de amostragem de $\bar{X}$ para $n = 5$ e $n = 25$ quando $\mu = 0$ e $\sigma = 1$.

```
n = 25; curve(dnorm(x, mean=0, sd=1/sqrt(n)), -3, 3,
              xlab='x',
              ylab='Desindades da média amostral', bty='l')
n = 5; curve(dnorm(x, mean=0, sd=1/sqrt(n)), add=TRUE)
n = 1; curve(dnorm(x, mean=0, sd=1/sqrt(n)), add=TRUE)
```



O Teorema Central do Limite

Conquanto o centro permanece o mesmo, a variabilidade de $\bar{X}$ fica menor à medida que n aumenta. Se o tamanho da amostra aumenta por um fator 4, o desvio-padrão será $\frac{1}{2}$ de suas populações. A densidade se concentrará na média. Isto é, com maiores e maiores probabilidades, o valor aleatório de $\bar{X}$ será mais perto da média, μ, da população parental. Esse fênomeno de concentração ao redor da média é conhecido como *lei dos grandes números*.

O Teorema Central do Limite

O *teorema central do limite* diz que para qualquer população parental com média μ e desvio-padrão σ, a distribuição de amostragem de $\bar{X}$ para n grande satisfaz

$$P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b) \approx P(Z \leq b),$$

onde Z é uma variável aleatória normal padrão. Isto é, para n grande o suficiente, a distribuição padronizada de $\bar{X}$ será aproxidamente uma normal padrão.

Inferência Estatística

Essa seção do nosso **Curso Formação Cientista de Dados** foca em conceitos básicos de inferência estatística usando simulação - evitando cálculos probabilísticos - de modo a produzir respostas. Como vimos, inferência estatística é o processo de extrair inferências sobre uma população com base em dados amostrais dessa população. Nesse processo, por suposto, são quatro os conceitos-chaves que devem ser levados em consideração:

- **populações** Nossas populações são modeladas por meio de distribuições que descrevem a aleatoriedade de cada ponto dos dados amostrais;
- **parâmetros** Um parâmetro é um número que descreve a população, tal como a média (μ) ou o desvio-padrão (σ);

Inferência Estatística

- **amostras** Uma amostra é uma coleção de observações da população, onde uma observação é a realização de uma variável aleatória com a distribuição da população. Para nossos propósitos, uma amostra é usualmente assumida como sendo uma amostra aleatória da população, o que implica em independência;
- **estatística** Um sumário numérico da amostra. Por exemplo, a média amostral $\bar{x}$ ou o desvio-padrão amostral s.

Inferência Estatística

A população é descrita pelos parâmetros e é representada pelas amostras. Dado que uma estatística sumariza a amostra, uma questão estatística central é como nós podemos inferir informação sobre os parâmetros a partir de uma estatística?

Ao comparar a densidade interna para a distribuição de $\bar{x}$ com a distribuição teórica de um valor individual, nós podemos ser levados a dizer que a distribuição de $\bar{x}$ está centrada no mesmo lugar que a da população - isto é, a aleatoriedade de uma estatística está centrada na aleatoriedade da população e a distribuição é bem comportada. Entretanto, é evidente que a distribuição de $\bar{x}$ tem uma variabilidade menor do que a da população. Em geral, para a simulação de $\bar{x}$ as seguintes observações podem ser explicitadas:

Inferência Estatística

- Para qualquer amostra aleatória, a distribuição amostral de $\bar{x}$ é centrada na média populacional, μ;
- Para qualquer amostra aleatória, a distribuição amostral de $\bar{x}$ possui um desvio-padrão de $\frac{\sigma}{\sqrt{n}}$ onde σ é o desvio padrão da população;
- Ademais, se a população for normalmente distribuída, a distribuição de $\bar{x}$ também será normalmente distribuída.

Inferência Estatística

O ponto é, dados simulados podem ser utilizados para investigar relacionamentos, de modo a produzir respostas que embora não sejam sempre precisas, podem dar excelentes *insights*. Essa seção, por suposto, cobre os conceitos básicos para performar simulações com o R e posteriormente aplica isso para introduzir a abordagem mais utilizada para construir inferência estatística.

Simulação

Para nossas simulações, o ponto inicial é especificar um modelo probabilístico para os dados. O R possui, como vimos, funções específicas para produzir amostras aleatórias a partir da distribuição da população. Nssas simulações usualmente envolvem sumarizar uma amostra aleatória com uma estatística. Por exemplo, encontrar uma realização da média amostral de uma amostra aleatória de tamanho 16 obtida de uma população normal.

Inferência Estatística

```
mu = 100; sigma = 16 # parâmetros populacionais
x = rnorm(16, mu, sigma) # nossa amostra
mean(x) # média amostral
```

```
## [1] 98.86523
```

Inferência Estatística

O modelo é especificado nesse exemplo pela escolha da família de distribuição (`rnorm`) e pela escolha dos parâmetros. Com isso, o R pode ser utilizado para produir uma amostra aleatória de um tamanho determinado. O valor *mean(x)* é o valor produzido. Ao rodar o código acima repetidas vezes, tudo o que você terá, a propósito, serão números diferentes. Isso traz *insights* interessantes sobre a forma da distribuição, seus mínimos, suas médias, variâncias, probabilidades relacionadas, etc.

Repetindo uma simulação

Repetir uma simulação, a propósito, é a chave para obter insights sobre essa simulação. Há diversas formas de repetir uma expressão no R. Vamos aqui mostrar isso através da função `for`.

```
mu = 100; sigma = 16
M = 10; n = 16
res = numeric(M)
for (i in 1:M){

  res[i] = mean(rnorm(n, mean=mu, sd=sigma))

}

res
```

```
##  [1] 104.03563 102.75109  93.70237 102.05196 100.36116  98.57477 104.51395
##  [8]  99.14903 107.93251  90.29412
```

Inferência Estatística

Essa é uma forma bastante direta, mas há outras formas. Por exemplo, podemos criar uma função para chamar nossa expressão a criar um único $\bar{x}$ para utilizar com `sapply`:

```
xbar = function(i)
  mean(rnorm(n, mean=mu, sd=sigma))

sapply(1:M, xbar)
```

```
##   [1] 104.21484 102.80604 105.88470  96.37395 103.01796  98.23108 103.81314
##   [8] 104.56875 103.49757  95.75595
```

Inferência Estatística

Ou simplesmente,

```
replicate(M, mean(rnorm(n, mean=mu, sd=sigma)))
```

```
##   [1] 100.27518  98.66045 108.15381 102.64012 105.21034 104.36471 112.54579
##   [8] 100.78001  99.25324  96.20105
```

Inferência Estatística

Com a função apply:

```
x = matrix(rnorm(M*n, mean=mu, sd=sigma), nrow=n)
dim(x)
```

```
## [1] 16 10
```

```
apply(x, 2, mean)
```

```
##  [1] 107.76388 102.66261 104.75218 100.91053 102.37037  99.29138 106.64731
##  [8]  97.65304 103.03606 101.28149
```

Testes de significância

Imagine agora um estudo de algum novo tratamento que melhora a performance. Por exemplo, consumir uma porção de mel durante os exercícios aumenta a performance? De modo a testar esse tipo de tratamento, uma coorte de sete pessoas convenientemente selecionadas é obtida. O pesquisador aleatoriamente atribui a três como grupo de controle e a 4 como grupo de tratamento. O grupo de controle deveria dar uma base de comparação para o grupo de tratamento. Os dados coletadas são alguma medida de modo que valores menores são melhores:

Inferência Estatística

```
controle = c(23, 33, 40)
tratamento = c(19, 22, 25, 26)
data = stack(list(controle=controle, tratamento=tratamento))
aggregate(values~ind, data, mean)
```

```
##          ind values
## 1   controle     32
## 2 tratamento     23
```

Inferência Estatística

Uma diferença de 9 é obtida. O pesquisador fica então emocionado. O estudo parece mostrar que uma simples porção de mel pode aumentar a performance. É preciso, contudo, ser cauteloso. Os resultados podem ser derivados simplesmente da aleatoriedade da escolha. Isto é, pessoas que performam melhor podem estar no grupo de tratamento, enquanto que pessoas que naturalmente performam pior pode estar no grupo de controle, de modo que a porção de mel não impactou em nada. A partir da simulação, podemos investigar se é esse o caso. Em particular, nós podemos enumerar todas as possibilidades de diferentes combinações. A função `combn` irá listar então através dos seus índices.

Inferência Estatística

```
cmbs = combn(7, 3) # 35 possibilidades
cmbs[,1:6]
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    1    1    1    1    1
## [2,]    2    2    2    2    2    3
## [3,]    3    4    5    6    7    4
```

Inferência Estatística

O primeiro seria apenas a aleatorização observada.

```
i = 1
ind = cmbs[,1]
obs = mean(data$values[ind] - mean(data$values[-ind]))
obs
```

```
## [1] 9
```

O uso da indexação negativa é conveniente nesse caso, dado que estamos parcionando nosso conjunto de 7 elementos em grupos de 3 e 4. De modo a computar as 35 combinações, podemos utilizar a função `apply`.

Inferência Estatística

```
res = apply(cmbs, 2, function(ind) {
    mean(data$values[ind]) - mean(data$values[-ind])
})
```

Os valores em `res` representam a distribuição para a diferença no grupo das médias. Agora que nós temos 35 diferentes valores, nós podemos ver o quão extremo o número 9 é nós podemos contar quantos valores são iguais ou maiores do que ele.

Inferência Estatística

```
sum(res >= obs)
```

```
## [1] 3
```

```
sum(res >= obs)/length(res)
```

```
## [1] 0.08571429
```

Inferência Estatística

Apenas três das 35 atribuições aleatórias ou simples 8.57% irão produzir uma diferença igual ou maior do que 9. Esse valor parece ser, portanto, improvável.

Inferência Estatística

Estimação e intervalos de confiança

A última subseção procurou verificar se um tratamento induzia melhor performance. Nessa, temos uma questão diferente. Como uma estatística amostral pode produzir uma boa estimativa de um parâmetro?

Uma notícia comum é um relatório sobre uma proporção em uma amostra. Por exemplo, em meados de 2012, seguindo as decisões da Suprema Corte norte-americana sobre a defesa do casamento gay, uma pesquisa de opinião foi feita pela Princeton Survey Research. Os pesquisadores perguntaram para uma amostra aleatória de 1.003 norte-americanos se o casamento gay deveria ser reconhecido como válido, tendo os mesmos atributos do casamento entre pessoas de sexos distintos. O resultado foi que 55% eram a favor da validade do casamento entre pessoas do mesmo sexo, uma resposta bastante elevada.

Inferência Estatística

O valor 55%, por suposto, representa a amostra, de modo que ele é uma estatística. A implicação é que de alguma maneira esse valor representa um parâmetro - a proporção de todos os norte-americanos. De que modo então essa estatística estima um parâmetro desconhecido?

Novamente, nós começamos com algum modelo probabilístico para os dados. No cenário acima, um modelo simples seria aquele em que cada pessoa escolhida aleatoriamente possui uma probabilidade p de responder sim. Isto é, as variáveis aleatórias possuem uma distribuição `Bernoulli(p)`. Outros cenários podem assumir uma distribuição normal para as variáveis aleatórias que produzem a amostra. Os modelos que consideramos são definidos em meio a uma família e um ou mais parâmetros.

Inferência Estatística

A questão de uma estimação estatística de um parâmetro é, então, respondida por meio da observação de valores tais como $\bar{x}$ em relação μ, ou $\hat{p}$ em relação a $\hat{p}$. Existem várias maneiras de comparar uma relação, uma simples é olhar para suas diferenças.

Vamos focar nas diferenças, por enquanto. O quanto podemos dizer? Se nós tivermos uma amostra grande, a intuição nos diria que a média amostral é uma estimativa melhor para μ do que um valor único. Por que? Ambas possuem um valor esperado - μ. O segredo é a variabilidade. Para um amostra aleatória de tamanho n, $VAR(\bar{x}) = \frac{\sigma^2}{n}$, onde σ^2 é a variância da população.

Inferência Estatística

Deixemos que θ seja algum parâmetro e que $\hat{\theta}$ seja alguma estatística para estimar θ. Olhar para a variabilidade, ou $E(\hat{\theta} - \theta)^2$, é razoável. Para uma variável aleatória, isso pode ser escrito de duas formas

$$E((\hat{\theta} - \theta)^2) = VAR(\hat{\theta}) + \left[E(\hat{\theta} - \theta)^2\right] = \quad \text{variância} + \text{viés}^2. \quad (2)$$

O viés é simplesmente a diferença entre o valor esperado do estimador e do parâmetro. A maioria das estatísticas que nós encontramos como não-viesada, significa que essa diferença é zero. Exemplos são, $E(\bar{x}) = \mu$ e $E(\hat{p}) = p$.

Inferência Estatística

Uma questão mais sutil do que olhar para a expectativa é olhar para a distribuição de amostragem. Para um modelo de população normal, nós vimos que nós podemos simular a distribuição de $\bar{x} - \mu$ da seguinte forma.

```
mu = 100; sigma = 16
M = 1000; n = 4
res = replicate(4, mean(rnorm(n, mu, sigma)) - mu)
```

Inferência Estatística

O desvio-padrão de $\bar{x}$, como vimos, é $\frac{\sigma}{\sqrt{nn}}$, de modo que se substituírmos σ por s, o que obteremos será o erro-padrão, isto é, o estimador substitui o parâmetro desconhecido. A distribuição depende de n, mas de posse de uma amostra grande o suficiente nós podemos fazer perguntas do tipo: onde está a maioria dos dados? Sendo preciso, qual intervalo de cerca de 0 contém 95% dos dados?

Inferência Estatística

Podemos simular a *estatística t* para responder essa pergunta:

```
mu = 100; sigma = 16
M = 1000; n = 4

res = replicate(M, {
  x = rnorm(n, mu, sigma)
  SE = sd(x)/sqrt(n)       # erro padrão
  (mean(x) - mu)/SE
})
```

Inferência Estatística

De modo a encontrar um intervalo, nós usamos a função `quantile`:

```
quantile(res, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## -2.985613  3.336426
```

Inferência Estatística

Então, basicamente, nós temos com probabilidade aproximada de 95% que

$$-3.05 * SE < \bar{x} - \mu < 3.19 * SE.$$

Ou, em outras palavras, o valor de $\bar{x}$ não é muitos erros-padrão distante da média, μ, com probabilidade elevada. O segredo dessa declaração é comparar distâncias usando uma escala definida pelo erro-padrão.

Inferência Estatística

Quando nós rodamos uma simulação, nós conhecemos parâmetros como μ a partir da especificação do nosso modelo e pela geração de valores repetidos de uma estatística como $\bar{x}$. Na vida real, existe uma perspectiva diferente: nós temos um único valor para a estatística e nós não conhecemos o parâmetro. O que nós podemos dizer sobre um parâmetro baseado em um único valor de $\bar{x}$?

Inferência Estatística

Nós podemos simplesmente inverter a algebra e resolver para μ de modo a dizer que com probabilidade 0.95,

$$\bar{x} - 3.19 * SE < \mu < \bar{x} + 3.05 * SE.$$

Essa fórmula está descrevendo a relação como variáveis aleatórias. Nós temos uma única realização, de modo que nós precisamos ser cautelosos com o termo *probabilidade*, de modo que a frase se torna *nós estamos 95% confiantes de que o intervalo* $\bar{x} - 3.19 * SE, \bar{x} + 3.05 * SE$ *contém o parâmetro desconhecido* μ.

G. Casella and R. L. Berger. *Inferência Estatística*. Cengage Learning, 2016.

P. L. Meyer. *Probabilidade - Aplicações á Estatística*. LTC, 2011.

J. Verzani. *Using R for Introductory Statistics*. CRC Press, 2014.