Resumo do Artigo “actuar: An R Package for Actuarial Science”

Gabriel D’assumpção de Carvalho

2024-12-09

Contents

1 Introdução

O objetivo deste documento é apresentar um resumo do segundo tópico do artigo **actuar: An R Package for Actuarial Science**. O artigo fornece uma explicação acessível e menos técnica sobre o pacote `actuar`, que disponibiliza ferramentas essenciais para a Ciência Atuarial.

O pacote `actuar` oferece funcionalidades que abrangem as seguintes áreas principais:

- **Modelagem de distribuições de perdas** — Permite a definição e ajuste de distribuições de perdas, que são fundamentais para o cálculo de prêmios e provisões técnicas.
- **Teoria do risco** — Facilita o estudo de processos estocásticos associados a sinistros e eventos de risco.
- **Simulação de modelos hierárquicos compostos** — Possibilita a criação de modelos de sinistros compostos, que são frequentemente utilizados para representar a variabilidade na frequência e severidade das perdas.
- **Teoria da credibilidade** — Suporta a aplicação de métodos de credibilidade, uma técnica que permite ajustar prêmios individuais com base em informações coletivas e individuais.

2 Momentos Centrais de uma Variável Aleatória

Um cientista atuarial possui diversas habilidades que podem ser aplicadas em várias tarefas. Entre elas, uma das mais importantes é a **modelagem de variáveis aleatórias** de interesse. Essas variáveis frequentemente representam conceitos como **reservas monetárias**, **distribuição de valores de sinistros** ou **determinação de preços de produtos ou serviços**. Assim, é essencial que o atuário tenha ferramentas para modelar essas variáveis de maneira eficiente.

Na estatística, um dos parâmetros utilizados para descrever a forma da distribuição de uma variável aleatória X é o **momento central**. O **momento central de ordem** k de uma variável X é dado por:

$$\mu_k = E[(X - \mu_X)^k] \tag{1}$$

Aqui, $\mu_X = E[X]$ é a média da variável X, e o momento central de ordem k corresponde ao valor médio da k-ésima potência do desvio de X em relação à sua média.

A partir dos momentos centrais, podemos inferir diversos parâmetros importantes sobre a distribuição de X, como:

- **Variância (momento central de ordem 2)**:

$$\mu_2 = E[(X - \mu_x)^2] = \sigma^2 \tag{2}$$

A variância mede a dispersão dos valores de X em torno da média.

- **Desvio Padrão**:

$$\sqrt{\mu_2} = \sqrt{E[(X - \mu_x)^2]} = \sqrt{\sigma^2} = \sigma \tag{3}$$

O desvio padrão é a raiz quadrada da variância e fornece uma medida de dispersão da mesma forma, mas em unidades da própria variável.

- **Coeficiente de Variação**:

$$\frac{\sigma}{\mu_x} \tag{4}$$

O coeficiente de variação é uma razão entre o desvio padrão e a média, o que permite comparar a dispersão entre distribuições com diferentes médias.

- **Assimetria (momento central de ordem 3)**:

$$\text{Assimetria} = \frac{\mu_3}{\sigma^3} \tag{5}$$

A assimetria indica a simetria da distribuição em torno de sua média. Um valor positivo sugere que a distribuição tem cauda à direita, enquanto um valor negativo indica cauda à esquerda.

- **Curtose (momento central de ordem 4)**:

$$\text{Curtose} = \frac{\mu_4}{\sigma^4} \tag{6}$$

A curtose mede a "altitude" das caudas da distribuição. Uma curtose alta indica caudas pesadas, enquanto uma curtose baixa sugere caudas leves.

Como podemos ver, os **momentos centrais** fornecem informações essenciais sobre a distribuição de uma variável aleatória, permitindo inferir características importantes para a modelagem atuarial.

2.1 Momentos limitados

Existem diversos serviços que uma seguradora pode oferecer aos seus clientes, sendo a franquia de produto um dos mais comuns. Nesse tipo de serviço, a seguradora é responsável por cobrir as perdas até um limite x do valor total do sinistro X. Assim, a perda efetiva para a seguradora será o mínimo entre X e x. Dessa forma, a inferência sobre a variável de interesse pode ser realizada com base no seu **momento limitado**, definido por:

$$E[(X \wedge x)] = E[min(X, x)] \tag{7}$$

3 Dados Agrupados

Em ciências atuariais, uma das estruturas de dados mais utilizadas são as tabelas organizadas no formato **intervalo-frequência**. Essas tabelas apresentam os dados agrupados em intervalos, representados por $(c_{j-1}, c_j]$, onde $j = 1, \ldots, r$ denota os intervalos e n_j corresponde à quantidade de eventos registrados no intervalo j.

4 Exemplo de Aplicação

A tabela a seguir mostra os casos de AIDS identificados no Brasil no ano de 2023, classificados por faixas etárias.

Faixa Etária	Frequência
(0, 4]	88
(4, 12]	40
(12, 19]	318
(19, 24]	1.578
(24, 29]	2.607
(29, 34]	2.481
(34, 39]	2.232
(39, 49]	3.693
(49, 59]	2.174
(59, 65]	1.070

Table 1: Casos de AIDS identificados no Brasil segundo faixa etária em 2023 - Datasus.

4.1 Observação

Foi considerada como idade máxima da população o limite de 65 anos.

4.2 Código R para Representação dos Dados

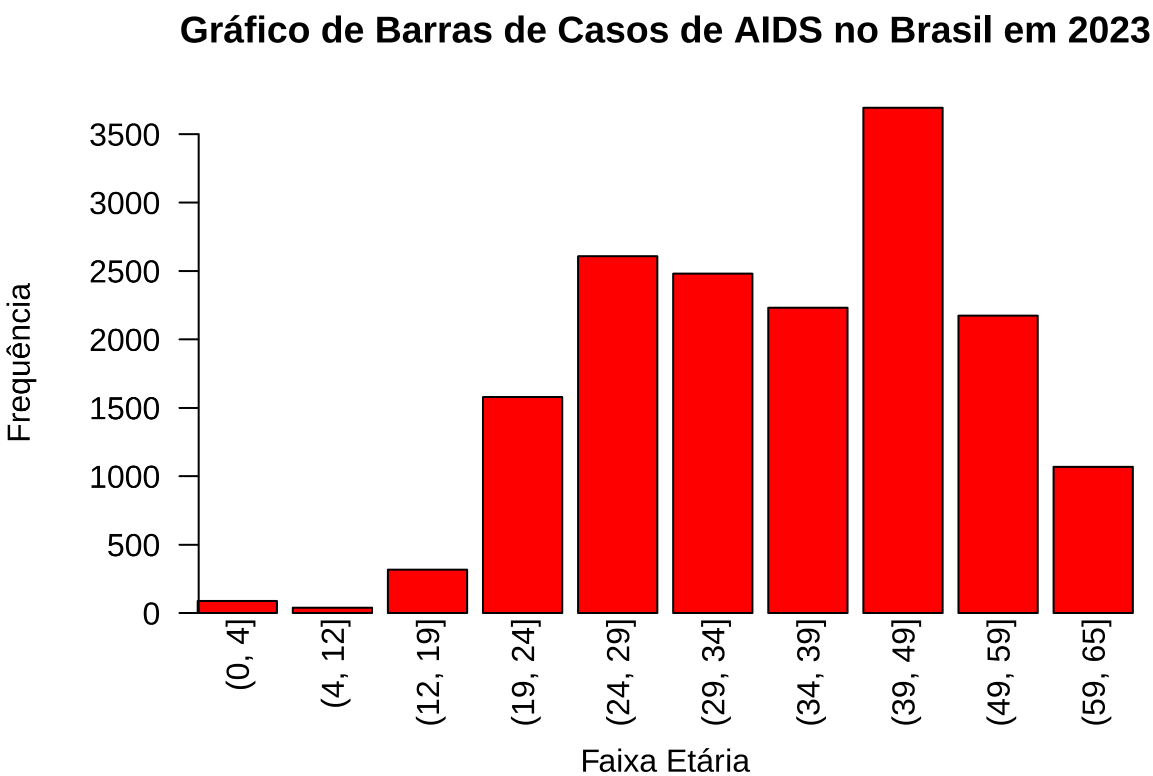
Abaixo está o código R utilizado para representar os dados agrupados:

```r
# Criação de dados agrupados
x <- grouped.data(Group = c(0, 4, 12, 19, 24, 29, 34, 39, 49, 59, 65),
                  aids = c(88, 40, 318, 1578, 2607, 2481, 2232, 3693, 2174, 1070))

# Impressão dos dados
print(x)
```

```
##       Group aids
## 1   (0,  4]   88
## 2   (4, 12]   40
```

```
## 3  (12, 19]  318
## 4  (19, 24] 1578
## 5  (24, 29] 2607
## 6  (29, 34] 2481
## 7  (34, 39] 2232
## 8  (39, 49] 3693
## 9  (49, 59] 2174
## 10 (59, 65] 1070
```

```r
# Criando o gráfico de barras
par(mgp = c(3.2, 0, -1)) # Aumenta o afastamento do título do eixo X
barplot(x[, 2],
        main = "Gráfico de Barras de Casos de AIDS no Brasil em 2023",
        xlab = "Faixa Etária",
        ylab = "Frequência",
        col = "red",
        border = "black",
        names.arg = c("(0, 4]", "(4, 12]", "(12, 19]", "(19, 24]",
                      "(24, 29]", "(29, 34]", "(34, 39]", "(39, 49]",
                      "(49, 59]", "(59, 65]"),
        las = 2)  # Para os rótulos do eixo X ficarem verticais
```



O gráfico de barras é uma excelente ferramenta para representar dados agrupados, especialmente quando lidamos com faixas de tempo. Ele facilita a visualização das tendências à medida que os dados avançam nas faixas etárias. Ao observar o gráfico de casos de AIDS no Brasil de 2023, é possível perceber que a faixa etária de (0, 19] apresenta as menores incidências, possivelmente devido à menor atividade sexual nessa faixa etária. A partir dessa faixa, as incidências começam a crescer. Notavelmente, a moda (a faixa com maior frequência) ocorre na faixa etária de (39, 49]. No entanto, é importante determinar o intervalo médio para uma análise mais precisa.

4.3 Cálculo do Intervalo Médio

O intervalo médio é calculado com base na fórmula:

$$\sum_{j=1}^{r}(\frac{c_{j-1}, c_j}{2})n_j \tag{8}$$

onde:

- c_{j-1} e c_j são os limites inferior e superior de cada intervalo, respectivamente.
- n_j é a frequência (número de eventos) registrada no intervalo j.

4.3.1 Implementação no R

O cálculo do intervalo médio é realizado pela função `mean_grouped` abaixo:

```r
mean_grouped = function(df) {
  total = 0
  for (i in 1:(length(df[, 1]) - 1)) {
    total = total + ((df[, 1][i] + df[, 1][i + 1]) / 2) * df[, 2][i]
  }
  return(total / sum(df[,2]))
}

# Calcular a média ponderada (intervalo médio)
print(mean_grouped(x))
```

```
## [1] 37.73018
```

```r
# Comparar com a média simples calculada pela função mean() do R
print(mean(x))
```

```
##     aids
## 37.73018
```

Como pode ser observado, o cálculo da faixa etária média obtido pela função `mean_grouped` resulta em 38,55 anos. Esse valor reflete a média ponderada das faixas etárias, considerando a distribuição dos casos em cada intervalo. A fórmula usada pela função é semelhante à fórmula da **esperança matemática** de uma distribuição uniforme, que para um intervalo (a, b) é dada por:

$$E[x] = \frac{b+a}{2} \tag{9}$$

onde a é o limite inferior e b o limite superior do intervalo. Essa fórmula é a base para o cálculo de cada ponto médio dos intervalos, o que faz com que o cálculo realizado pela função `mean_grouped` seja uma aproximação eficiente da média ponderada da distribuição das faixas etárias.

4.4 Função Empírica de Distribuição Acumulada (CDF)

A **CDF** (Cumulative Distribution Function) $F(x)$ representa a probabilidade de que a variável aleatória X assuma um valor menor ou igual a x, ou seja, $P(X \leq x)$. O seu complementar é a **função de sobrevivência** $S(x)$, que é definida como:

$$S(x) = 1 - P(X \leq x) = P(X \geq x) \tag{10}$$

4.5 Frequência Agrupada e Fórmula de $\hat{F}_n(x)$

A **frequência agrupada** é uma forma de organizar dados em intervalos, onde as frequências indicam o número de observações em cada intervalo. Para calcular a **função empírica de distribuição acumulada** $\hat{F}_n(x)$ em dados agrupados, utilizamos a seguinte fórmula:

$$\hat{F}_n(x) = \begin{cases} 0, & x \leq c_0 \\ \frac{(c_j - x)F_n(c_{j-1}) + (x - c_{j-1})F_n(c_j)}{c_j - c_{j-1}}, & c_{j-1} < x \leq c_j \\ 1, & x > c_r \end{cases} \tag{11}$$

Onde:

- $c_0, c_1, \ldots, c_r$ são os limites dos intervalos;
- $F_n(c_j)$ é o valor da função acumulada no limite superior do intervalo c_j;
- $F_n(c_{j-1})$ é o valor da função acumulada no limite inferior do intervalo c_{j-1};
- $c_j - c_{j-1}$ é a amplitude do intervalo;
- x é o ponto no qual se deseja calcular $\hat{F}_n(x)$.

4.6 Como Calcular $F_n(c_j)$

Para calcular $F_n(c_j)$, utilizamos a seguinte fórmula:

$$F_n(c_j) = \frac{\sum_{i=1}^{j} n_i}{N} \tag{12}$$

Onde:

- n_i é a frequência do i-ésimo intervalo;
- $N = \sum_{i=1}^{r} n_i$ é o total de observações.

4.6.1 Implementação no R

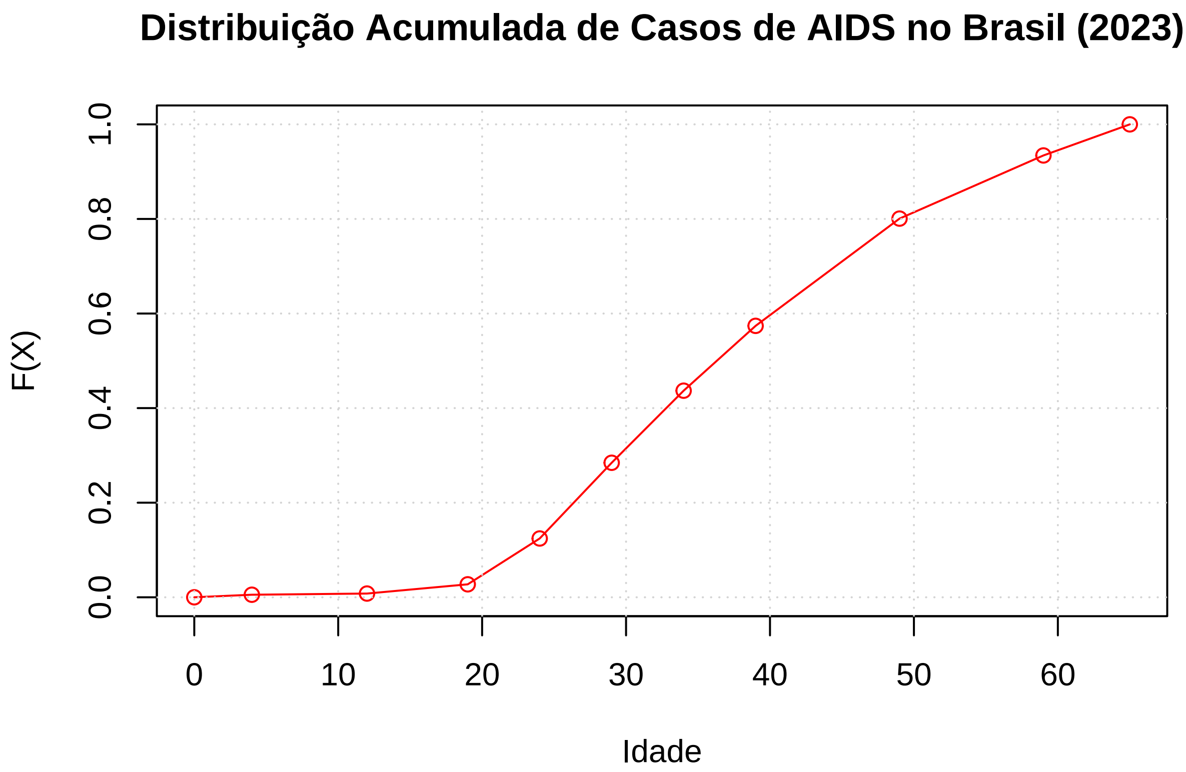
O código abaixo implementa a função que calcula a função de distribuição acumulada (ogiva) para dados agrupados.

```r
def_ogive = function(df) {
  F_x = numeric()
  N = sum(df[, 2])
  for (i in 0:length(df[, 2])) {
    F_ci = sum(df[, 2][0:i]) / N
    F_x = c(F_x, F_ci)
  }
  return(F_x)
}

ogive_values = def_ogive(x)
```

4.6.2 Gráfico da Função de Distribuição Acumulada

```r
# Criar o gráfico de linhas
plot(rep(x[, 1]), ogive_values,
     type = "o",                # Tipo "o" conecta os pontos com linhas e marca os pontos
     col = "red",              # Cor da linha
     xlab = "Idade",            # Rótulo do eixo x
     ylab = "F(X)",             # Rótulo do eixo y
```

```r
     main = "Distribuição Acumulada de Casos de AIDS no Brasil (2023)") # Título do gráfico

# Adicionar uma grade para facilitar a leitura
grid()
```

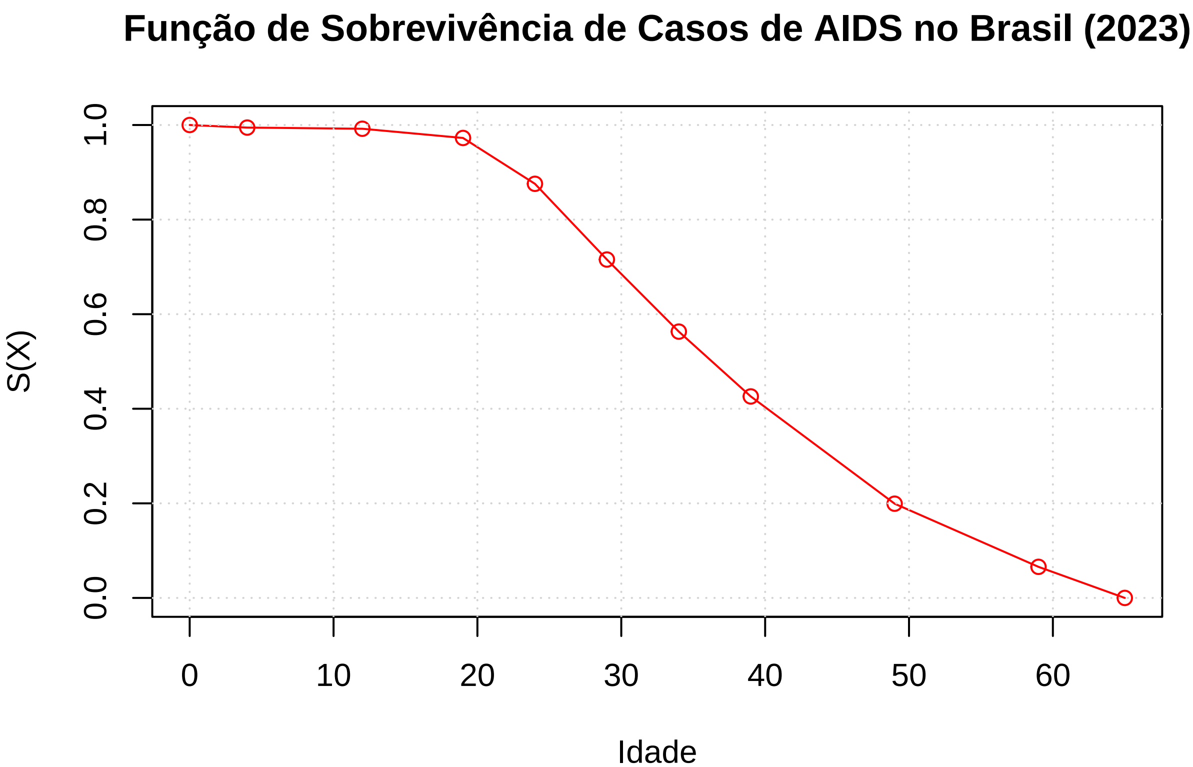


Podemos observar graficamente que o valor de $F(X) = 0.5$ ocorre entre as idades de 35 a 40 anos. Isso está de acordo com o cálculo da média, que foi de aproximadamente 38 anos. Esse comportamento reflete o fato de que cerca de 50% dos casos de AIDS estão em pessoas com idades inferiores a esse intervalo.

4.6.3 Gráfico da Função de Sobrevivência

Para calcular a função de sobrevivência, utilizamos a relação $S(x) = 1 - F(x)$. O código a seguir gera o gráfico de $S(x)$ para os casos de AIDS no Brasil em 2023.

```r
# Criar o gráfico de linhas
plot(rep(x[, 1]), 1 - ogive_values,
     type = "o",                   # Tipo "o" conecta os pontos com linhas e marca os pontos
     col = "red",                 # Cor da linha
     xlab = "Idade",               # Rótulo do eixo x
     ylab = "S(X)",                # Rótulo do eixo y
     main = "Função de Sobrevivência de Casos de AIDS no Brasil (2023)") # Título do gráfico

# Adicionar uma grade para facilitar a leitura
grid()
```



O gráfico da **função de sobrevivência** $S(x)$ reflete a proporção de casos de AIDS para as faixas etárias **acima de** x. Como $S(x)$ é o complementar de $F(x)$, ele mostra a **probabilidade de um caso ocorrer em uma faixa etária superior** àquela representada no eixo x.

Ao observar o gráfico de $S(x)$, notamos que ele é **decrescente**, o que indica que, à medida que a idade aumenta, a proporção de casos restantes (ou "sobreviventes") diminui. Em outras palavras, a cada faixa etária superior, o número de casos de AIDS vai ficando progressivamente menor, o que é esperado considerando a estrutura dos dados.

Vale ressaltar que, devido à forma como os dados foram estruturados (onde todos os indivíduos foram diagnosticados com AIDS), a **função de sobrevivência** neste contexto não reflete uma taxa de sobrevivência real, como seria o caso em um estudo de tempo até a morte ou cura de uma doença.

4.7 K-ésimo Momento Empírico

O cálculo do k-ésimo momento empírico para dados agrupados é realizado de forma diferente em relação aos dados não agrupados. Para dados agrupados, utilizamos a seguinte fórmula:

$$\hat{\mu}_k = \frac{1}{n} \sum_{j=1}^{r} \frac{n_j (c_j^{k+1} - c_{j-1}^{k+1})}{(k+1)(c_j - c_{j-1})} \tag{13}$$

Onde:

- $\hat{\mu}_k$: k-ésimo momento empírico.
- n_j: frequência do j-ésimo intervalo.
- c_{j-1} e c_j: limites inferior e superior do j-ésimo intervalo, respectivamente.
- r: número total de intervalos.
- n: número total de observações, dado por $n = \sum_{j=1}^{r} n_j$.

A principal diferença em relação ao cálculo do momento empírico de dados individuais é que, para dados agrupados, precisamos considerar a contribuição de cada intervalo com base nos seus limites e na frequência de observações n_j.

4.7.1 Implementação no R

A implementação a seguir permite calcular o k-ésimo momento empírico para uma tabela de frequências agrupadas. Veja o código abaixo:

```
empirical_moment = function(df, k){
  N = sum(df[,2])
  total = 0
  for (i in 1:length(df[,2])){
    total = total + (df[,2][i] * (df[,1][i+1] ** (k + 1) - df[,1][i] ** (k+ 1))) / ((k + 1) * (df[,1][i
  }
  return(total/N)
}

# Comprovante que o primeiro momento é igual a média

mean_x = empirical_moment(x, 1)
print(mean_x)
```

```
## [1] 37.73018
```

```
print(mean(x))
```

```
##     aids
## 37.73018
```

```
# Cálculo Variância
var_x = empirical_moment(x, 2) - mean_x ** 2
print(var_x)
```

```
## [1] 157.9854
```

```
# Comparativo com função emm() do pacote actuar
moments = numeric(4)
for (i in 1:4) {
  moments[i] = empirical_moment(x, i)
}
print(moments)
```

```
## [1] 3.773018e+01 1.581552e+03 7.202931e+04 3.503155e+06
```

```
print(emm(x, order = 1:4))
```

```
## [1] 3.773018e+01 1.581552e+03 7.202931e+04 3.503155e+06
```