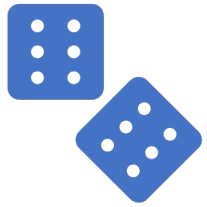
Aula 12 - Probabilidade e Distribuições

Prof. Dr. José Weligton Félix Gomes



Probabilidade e Distribuições

Amostragem Aleatória

- Muitos dos primeiros trabalhos em teoria da probabilidade foram sobre jogos e questões de jogo, com base em considerações de simetria.
- A noção básica, então, é a de uma amostra aleatória: lidar com um baralho de cartas bem embaralhado ou pegar bolas numeradas de uma urna bem mexida.
- Em R, você pode simular essas situações com a função de amostra. Se você quiser escolher cinco números aleatoriamente do conjunto 1:40, pode escrever ***sample*(*x*, *size* = n, *replace* = *FALSE*)**. Por exemplo:

> sample(1:40,5)

> sample(1:40,5, replace = FALSE ou TRUE)

- O primeiro argumento (x) é um vetor de valores a serem amostrados e o segundo (tamanho) é o tamanho da amostra. **replace** é onde você indica se a amostra deve ser feita **com reposição** (TRUE) ou **sem reposição** (**FALSE**). O padrão da amostra é amostragem sem reposição (FALSE).

Probabilidade e Distribuições

- Ou seja, **as amostras não conterão o mesmo número duas vezes** e o tamanho obviamente não pode ser maior do que o comprimento do vetor a ser amostrado. Se você deseja amostrar com reposição, então você precisa adicionar o argumento **replace = TRUE**.
- A amostragem com substituição é adequada para modelar lançamentos de moeda ou dados.
- Assim, por exemplo, para simular 10 lançamentos de moeda, poderíamos escrever

> sample(c ("H", "T"), 10, replace = T) #H = Heads e T = Tails.

- No lançamento de moeda justa, a probabilidade de cara deve ser igual à probabilidade de coroa, mas a ideia de um evento aleatório não se restringe a casos simétricos. Poderia ser igualmente bem aplicado a outros casos, como o resultado bem-sucedido de um procedimento cirúrgico. Esperançosamente, haveria uma chance melhor do que 50% disso.
- Você pode **simular dados com probabilidades nada iguais para os resultados** (digamos, 90% de chance de sucesso) usando o argumento prob para amostrar, como em

> sample(c("succ", "fail"), 10, replace=T, prob=c(0.9, 0.1))

Probabilidade e Distribuições

Cálculos de probabilidade e combinatória

- Ainda com relação ao caso de amostra sem reposição. Supondo que a probabilidade de se obter um determinado número como o primeiro da amostra deve ser de 1/40, o próximo de 1/39 e assim por diante.
- A probabilidade **de uma determinada amostra** deve ser 1 / (40 × 39 × 38 × 37 × 36).
- Em R, use a **função prod**, que calcula o produto de um vetor de números.

> 1/prod(40:36)

Probabilidade e Distribuições

Cálculos de probabilidade e combinatória

- No entanto, observe que essa é a probabilidade de obter determinados números em uma determinada ordem.
- Se fosse um jogo parecido com a loteria, você preferiria estar interessado na probabilidade de adivinhar um determinado conjunto de cinco números corretamente.
- Portanto, você também precisa incluir os casos que fornecem os mesmos números em uma ordem diferente.
- Visto que obviamente a probabilidade de cada um desses casos será a mesma, tudo o que precisamos fazer é descobrir quantos desses casos existem e multiplicar por isso.
- Existem cinco possibilidades para o primeiro número, e para cada uma delas existem quatro possibilidades para o segundo e assim por diante; ou seja, o número é 5 × 4 × 3 × 2 × 1.
- Este número também é escrito como 5! (5 fatorial). Portanto, a probabilidade de um "cupom de loteria ganhar" seria:

> prod(5:1)/prod(40:36)

Probabilidade e Distribuições

Cálculos de probabilidade e combinatória

- Existe outra maneira de chegar ao mesmo resultado. Observe que, uma vez que o conjunto real de números é imaterial, todos os conjuntos de cinco números devem ter a mesma probabilidade.
- Portanto, tudo o que precisamos fazer é calcular o número de maneiras de escolher 5 números entre 40. Isso é:
- $\binom{40}{5} = \frac{40!}{5!35!} = = 658008$
- Em R, a função **choose()** pode ser usada para calcular este número, e a probabilidade é, portanto,

> 1/choose(40,5)

Probabilidade e Distribuições

Distribuição Discreta

- Ao observar as replicações independentes de um experimento binário, você normalmente não estaria interessado em saber se cada caso é um sucesso ou um fracasso, mas sim no número total de sucessos (ou fracassos).
- Obviamente, esse número é aleatório, pois depende dos resultados aleatórios individuais e, consequentemente, é chamado de variável aleatória. Nesse caso, é uma variável aleatória de valor discreto que pode assumir valores $0, 1, \ldots, n$, onde n é o número de replicações.
- Uma variável aleatória X tem uma distribuição de probabilidade que pode ser descrita usando **probabilidades pontuais** $f\,(x) = P\,(X = x)$ ou a **função de distribuição cumulativa** $F\,(x) = P\,(X \leq x)$.
- No caso em questão, a distribuição pode ser calculada como tendo as probabilidades pontuais

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- Conhecida como **distribuição binomial**.

Probabilidade e Distribuições

Distribuição Contínua

- Alguns dados surgem de medições em uma escala essencialmente contínua, por exemplo, temperatura, concentrações, etc.
- Na prática, eles serão registrados com uma precisão finita.
- Para modelar dados contínuos, precisamos definir variáveis aleatórias que podem obter o valor de qualquer número real.
- Como há infinitos números infinitamente próximos, a probabilidade de qualquer valor particular será zero, então não existe probabilidade de ponto como para variáveis aleatórias de valor discreto.
- A função de distribuição cumulativa pode ser definida como antes, e temos a relação

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Probabilidade e Distribuições

Exemplos:

- A distribuição uniforme tem uma densidade constante em um intervalo especificado (por padrão [0, 1]).
- A distribuição normal (também conhecida como distribuição Gaussiana) tem densidade

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

Probabilidade e Distribuições

As distribuições integradas em R:

- Dois itens fundamentais podem ser calculados para uma distribuição estatística:
  - **Densidade ou probabilidade de ponto**:
    - A densidade para uma distribuição contínua é uma medida da probabilidade relativa de "obter um valor próximo de x". A probabilidade de obter um valor em um determinado intervalo é a área sob a parte correspondente da curva.
    - Para distribuições discretas, o termo "densidade" é usado para a probabilidade de ponto - a probabilidade de obter exatamente o valor x.

```
> x <- seq(-4,4,0.1)
> plot(x,dnorm(x),type="l") , onde "l" ➔ linha
> curve(dnorm(x), from=-4, to=4)
```

Probabilidade e Distribuições

As distribuições integradas em R:

- Probabilidade acumulada, função de distribuição:
- A função de distribuição cumulativa descreve a probabilidade de "acertar" x ou menos em uma determinada distribuição. As funções R correspondentes começam com um 'p' (para probabilidade) por convenção.
- Digamos que é sabido que alguma medida bioquímica em indivíduos saudáveis é bem descrita por uma distribuição normal com uma média de 132 e um desvio padrão de 13.
- Então, se um paciente tem um valor de 160, há

> **1-pnorm(160, mean=132, sd=13)**

[1] 0.01562612

- **Ou seja, apenas cerca de 1,5% da população geral, que tem esse valor ou superior.**
- A **função pnorm** retorna a probabilidade de obter um valor menor do que seu primeiro argumento em uma distribuição normal com a média e o desvio padrão fornecidos.

Probabilidade e Distribuições

As distribuições integradas em R:

- Para muitas pessoas, soa como uma contradição em termos de gerar números aleatórios em um computador, uma vez que seus resultados devem ser previsíveis e reproduzíveis.
- O que é possível, de fato, é gerar sequências de números “pseudo-aleatórios”, que para fins práticos se comportam como se tivessem sido sorteados ao acaso.
- O uso das funções que geram números aleatórios é direto. Por exemplo,

> **rnorm(10)** #Extrai 10 valores aleatórios de uma distribuição normal padrão.

> **rnorm (10, mean = 5, sd = 2)** #Extrai 10 valores aleatórios de uma distribuição normal com média 5 e desvio padrão 2.

Estatística Descritiva e Gráficos – Cap. 04

- É fácil calcular estatísticas descritivas com o R.
- **Por exemplo:** Média, Desvio Padrão, Variância, Mediana, quantis.
- **Funções no R**: mean( ), sd( ), var( ), median( ), quantile( ).
- Gráficos no R:
  - Histograma: hist( )
- Um propósito de calcular a função de distribuição cumulativa empírica (c.d.f.) é ver se os dados podem ser considerados normalmente distribuídos. Para uma melhor avaliação, você pode representar graficamente a k-ésima observação menor em relação ao valor esperado da k-ésima observação mais pequena de n em uma distribuição normal padrão.
  - Q-Q plot: qqnorm( )

Estatística Descritiva e Gráficos – Cap. 04

- Um “boxplot é um resumo gráfico de uma distribuição.
  - Box plot: boxplot( )

- Resumo Estatístico por grupo:
  - Ao lidar com dados agrupados, você frequentemente desejará ter várias estatísticas de resumo computadas dentro dos grupos.
  - A função **tapply** pega uma variável, divide-a de acordo com uma segunda variável e calcula a média para cada grupo. Da mesma forma, os desvios padrão e o número de observações nos grupos podem ser calculados.

Gráfico para dados Agrupados

- Ao lidar com dados agrupados, é importante ser capaz não apenas de criar plotagens para cada grupo, mas também de comparar as plotagens entre grupos.
- Algumas técnicas gráficas gerais que nos permitem exibir gráficos semelhantes para vários grupos na mesma página.
- Algumas funções têm recursos específicos para exibir dados de mais de um grupo.
  - Histograma – hist( )
  - Gráfico de Caixa – Boxplot
  - Tabelas – table( )