
FIAP

AULA 6

"Capacitar o aluno na construção de protótipos de Data Marts e Data Warehouse, utilizando ETL e a evolução para Data Ingestion."

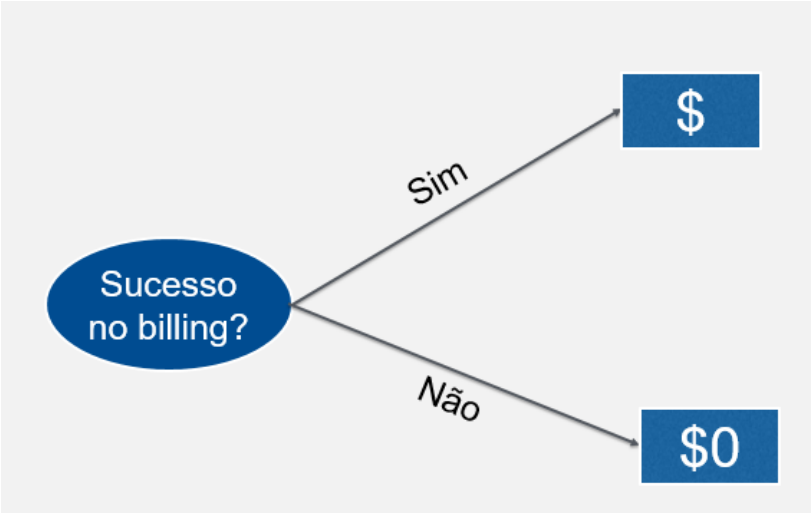
"Tornar o aluno capacitado a compreender e construir modelos estatísticos/analíticos básicos de classificação e preditivos utilizados em Data Science/Machine Learning."

Enterprise Analytics e Data Warehousing

Alcides C. Araújo

Problema

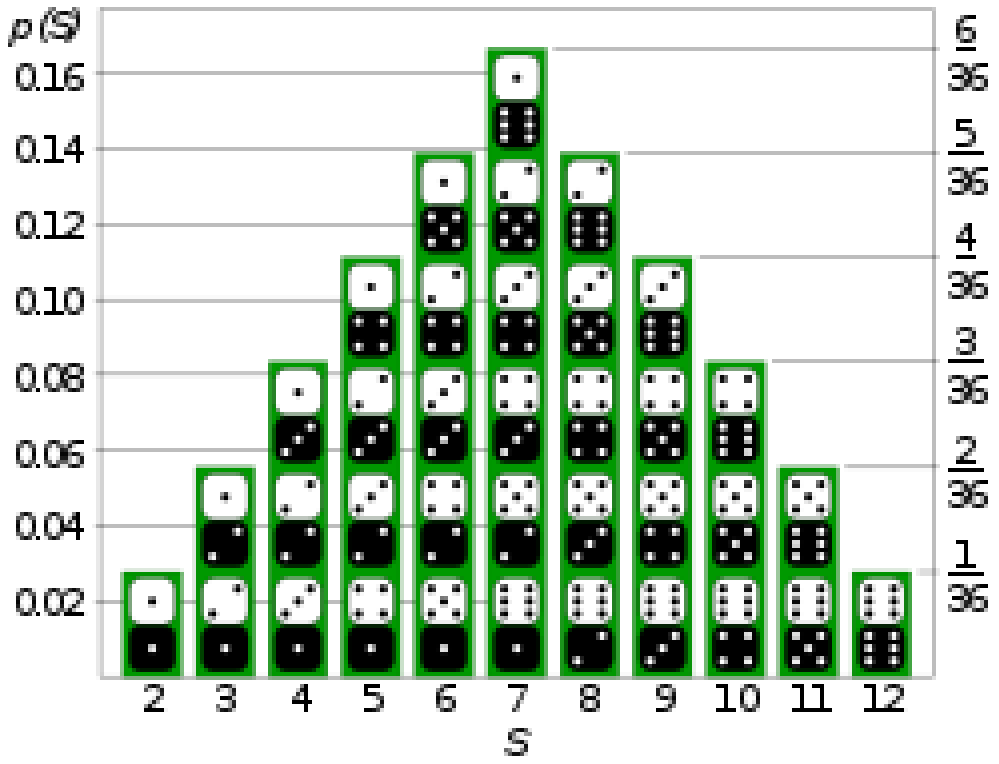
- Imagine este problema clássico de cobrança:



- Qual a probabilidade de sucesso?
- Se eu tenho uma base de 5 pessoas para serem cobradas. Qual a chance de cobrar entre 2 a 4 pessoas com sucesso?

Conceitos iniciais

- Distribuições de Probabilidade

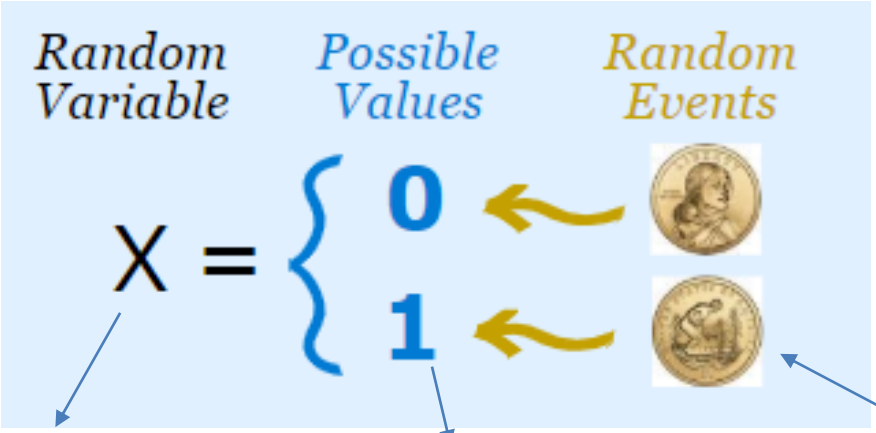


Conceitos

- Variável aleatória
- Eventos independentes
- Distribuições de probabilidade

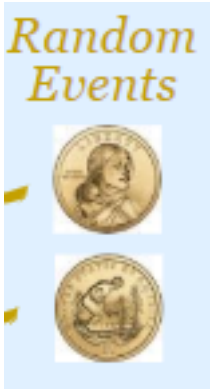
Variável aleatória

- Variáveis aleatórias representam resultados, em números, de processos aleatórios. Por exemplo, jogar uma moeda, prever o resultado de um jogo, tentar cobrar alguém, alugar um carro e não ter acidente.

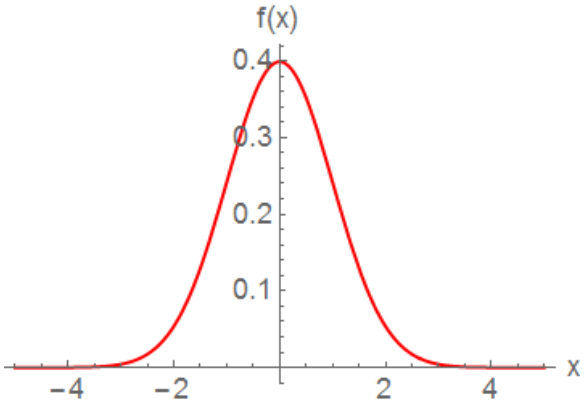


Tipos de variáveis aleatórias

- Discretas:
- Possui um número finito de resultados.
- Exemplo: Vivo/morto, grávida/não grávida, sucesso/fracasso, vitória/derrota/empate.



- Contínuas:
- Possui uma grande amplitude de resultados, em que se torna praticamente impossível contar o número de resultados possíveis.
- Exemplo: Peso, altura, renda, velocidade de um carro



Eventos independentes

- Eventos são independentes se o conhecimento de um evento não interfere na probabilidade de outro evento ocorrer.

- Imagine dois eventos A e B, a probabilidade de B ocorrer não é mudada, mesmo você sabendo que A ocorreu anteriormente.

Exemplo

Renda anual	Universidade A	Universidade B	TOTAL
menos 20k	36	24	60
20k - 39k	109	56	165
40k ou mais	35	40	75
TOTAL	180	120	300

Vamos calcular o seguinte:

P(Renda anual | Universidade) e P(Renda Anual).

Se P(Renda anual | Universidade) = P(Renda Anual) temos eventos independentes.

Lembrando que $P(B|A) = P(A \cap B)/P(A)$

P(menos 20k)	0,20
P(20k - 39k)	0,55
P(40k ou mais)	0,25

P(menos 20k | U. A)	0,20
P(20k - 39k | U. A)	0,61
P(40k ou mais | U. A)	0,19

P(menos 20k | U. B)	0,20
P(20k - 39k | U. B)	0,47
P(40k ou mais | U. B)	0,33

Vemos que P(menos 20k) = P(menos 20k | U. A) = P(menos 20k | U. B).
Ou seja, receber menos de 20k anuais é um evento independente de estudar na Universidade A ou B.

Exemplo detalhado

- Vamos estudar o caso de receber menos de 20k.

- $P(menos\ 20k) = \frac{60}{300} = 0{,}2$
- $P(menos\ 20k \mid U.A) = \frac{36/300}{180/300} = 0{,}2$
- $P(menos\ 20k \mid U.B) = \frac{24/300}{120/300} = 0{,}2$

Eventos independentes

- Agora veja o seguinte:

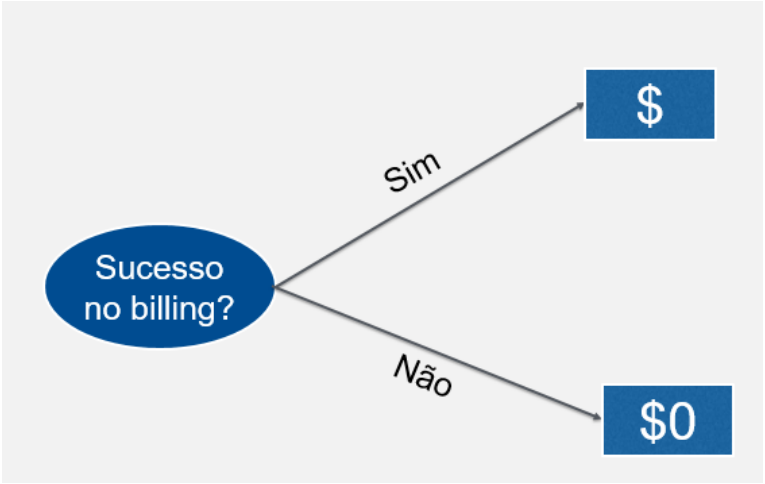
- Se $P(menos\ 20k \mid U.A) = \frac{P(menos\ 20k\ \cap U.A)}{P(U.A)}$ então:
- $P(menos\ 20k\ \cap U.\ A) = P(menos\ 20k \mid U.A) \cdot P(U.A)$
- Calculamos $P(menos\ 20k \mid U.A) = 0{,}2$, porém $P(menos\ 20k) = 0{,}2$ também.

- Quando isto acontece podemos generalizar que:
- $P(menos\ 20k\ \cap U.\ A) = P(menos\ 20k) \cdot P(U.A)$

- **Assim, a probabilidade de eventos independentes é o produto (multiplicação) de suas probabilidades.**

Relação entre probabilidade e distribuições

- Vamos lembrar do problema de cobrar 5 pessoas.



Distribuições de probabilidade

- Se atribuirmos como 1 para sucesso na cobrança e 0 para fracasso, podemos fazer a seguinte tabela:

Pessoa A	Pessoa B	Pessoa C	Pessoa D	Pessoa E
0	0	0	0	0
1	0	0	0	0
0	1	1	0	0
0	1	1	1	0
0	0	0	0	1

- Temos um problema de permutação ou combinação?
- A probabilidade de cobrança com sucesso da pessoa A depende da B ou C?

Combinações

- Baseado nas respostas anteriores, podemos construir o seguinte:

Número de sucessos	Combinações possíveis
0	1
1	5
2	10
3	10
4	5
5	1

Combinação: $C_k^n = \frac{n!}{(n-k)!k!}$

Para 0 sucessos: $C_0^5 = \frac{5!}{(5-0)!0!}$

Para 1 sucesso: $C_1^5 = \frac{5!}{(5-1)!1!}$

E assim por diante...

Probabilidades

- Observando o passado, temos 80% de cobrar com sucesso uma pessoa, ou seja, P(1) = 0,8 e P(0) = 0,2.
- Como estes eventos são independentes podemos construir o seguinte exemplo:

Pessoa A	Pessoa B	Pessoa C	Pessoa D	Pessoa E
0	0	0	0	0
1	0	0	0	0
1	1	0	0	0
1	1	1	0	0
1	1	1	1	0
1	1	1	1	1

- Linha 1: P(0 sucessos) = P(0) * P(0) * P(0) * P(0) * P(0) = $0{,}2^5$
- Linha 2: P(1 sucesso) = P(1) * P(0) * P(0) * P(0) * P(0) = $0{,}8^1 \cdot 0{,}2^4$
- Linha 3: P(2 sucessos) = P(1) * P(1) * P(0) * P(0) * P(0) = $0{,}8^2 \cdot 0{,}2^3$
- Linha 4: P(3 sucessos) = P(1) * P(1) * P(1) * P(0) * P(0) = $0{,}8^3 \cdot 0{,}2^2$
- Linha 5: P(4 sucessos) = P(1) * P(1) * P(1) * P(1) * P(0) = $0{,}8^4 \cdot 0{,}2^1$
- Linha 6: P(5 sucessos) = P(1) * P(1) * P(1) * P(1) * P(1) = $0{,}8^5$

a probabilidade de eventos independentes é o produto (multiplicação) de suas probabilidades.

Podemos generalizar para:

$$p^k \cdot (1-p)^{n-k}$$

Chamamos de probabilidade de Bernoulli

Juntar combinação e probabilidade

- Agora vamos criar uma tabela para os dois exemplos anteriores.

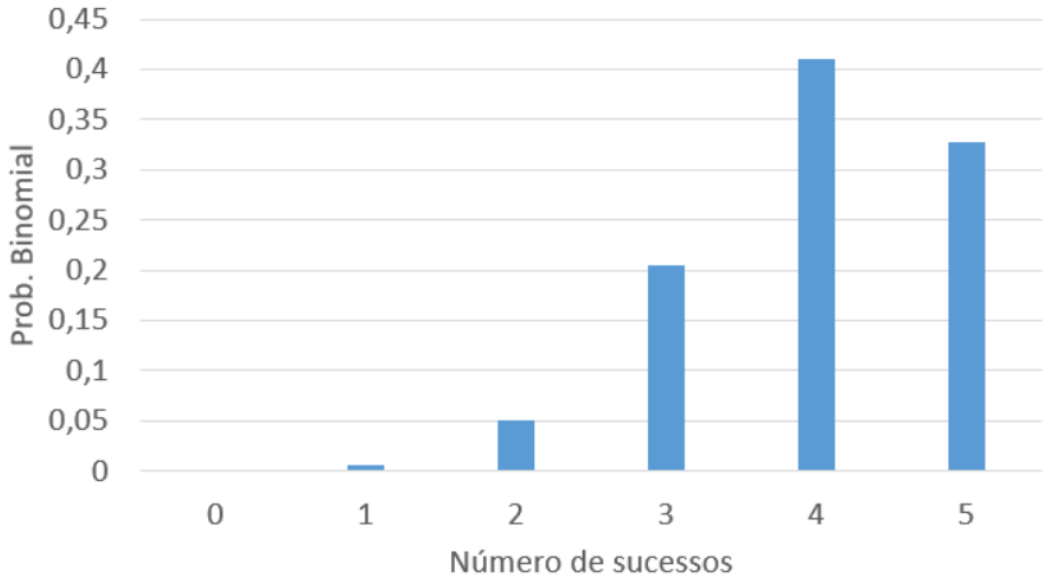
Número de sucessos	Combinações possíveis	Probabilidade de Bernoulli	Probabilidade Binomial
0	1	$0{,}2^5 = 0{,}00032$	$1 \cdot 0{,}2^5 = 0{,}00032$
1	5	$0{,}8^1 \cdot 0{,}2^4 = 0{,}00128$	$5 \cdot 0{,}8^1 \cdot 0{,}2^4 = 0{,}0064$
2	10	$0{,}8^2 \cdot 0{,}2^3 = 0{,}00512$	$10 \cdot 0{,}8^2 \cdot 0{,}2^3 = 0{,}0512$
3	10	$0{,}8^3 \cdot 0{,}2^2 = 0{,}02048$	$10 \cdot 0{,}8^3 \cdot 0{,}2^2 = 0{,}2048$
4	5	$0{,}8^4 \cdot 0{,}2^1 = 0{,}08192$	$5 \cdot 0{,}8^4 \cdot 0{,}2^1 = 0{,}4096$
5	1	$0{,}8^5 = 0{,}3277$	$1 \cdot 0{,}8^5 = 0{,}3277$

Nesta tabela temos um conjunto de resultados de eventos associados as suas respectivas probabilidades. Ou seja, denominamos isto de **Distribuição de Probabilidades**.

A probabilidade binomial é o resultado de uma sequencia de eventos em que somente 2 resultados podem ocorrer. Neste exemplo, sucesso ou fracasso.

Juntar combinação e probabilidade

- Podemos fazer um gráfico:

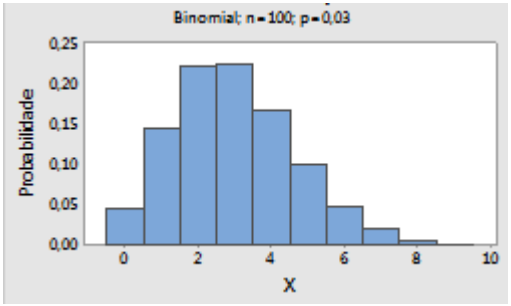


Gráficos deste tipo possuem um nome específico denominados de **Histogramas de frequência**.

Histogramas são gráficos de distribuições de probabilidades.

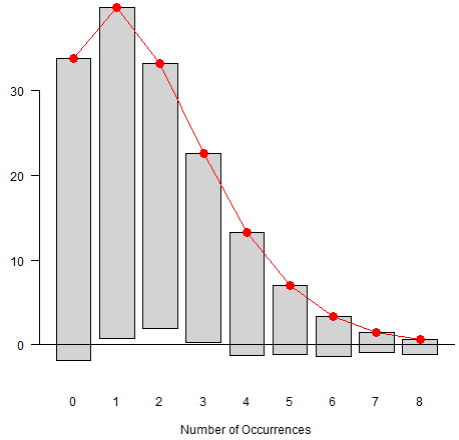
Distribuições de probabilidades

- Existem diversos padrões observados nestes gráficos.
- Os mais comuns são:
- Fiquem tranquilos que veremos isto no próximo semestre.



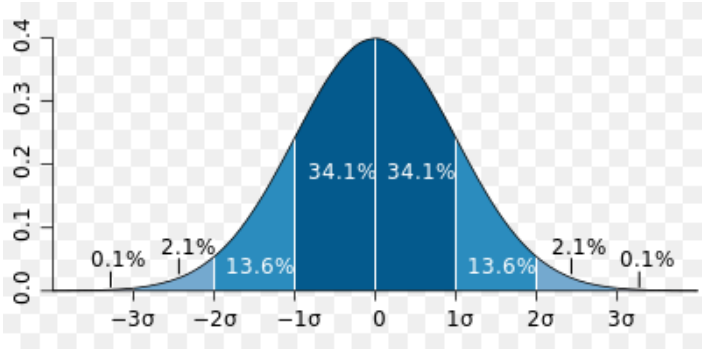
Distribuição binomial

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



Distribuição de Poisson

$$P(X = k) = \frac{e^{-\lambda} . \lambda^k}{k!}$$

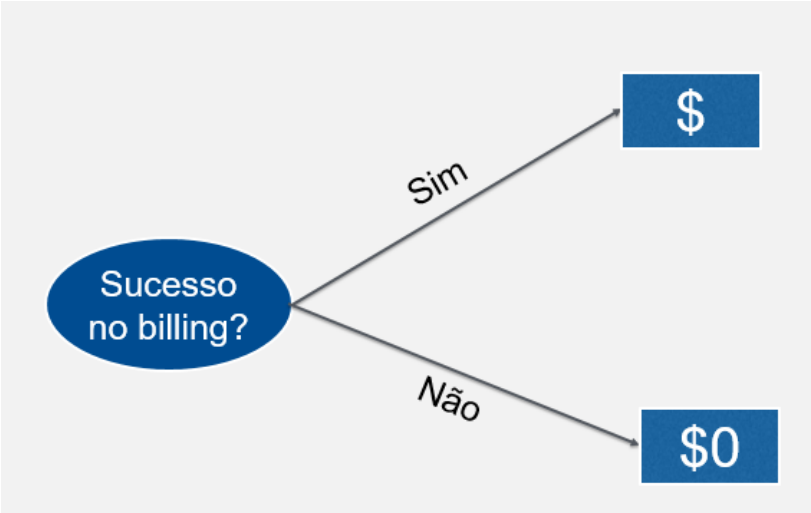


Distribuição Normal

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

Voltando ao nosso problema

- Imagine este problema clássico de cobrança:



- Qual a probabilidade de sucesso?
- Se eu tenho uma base de 5 pessoas para serem cobradas. Qual a chance de cobrar entre 2 a 4 pessoas com sucesso?

Resolução 1

- Qual a probabilidade de sucesso?

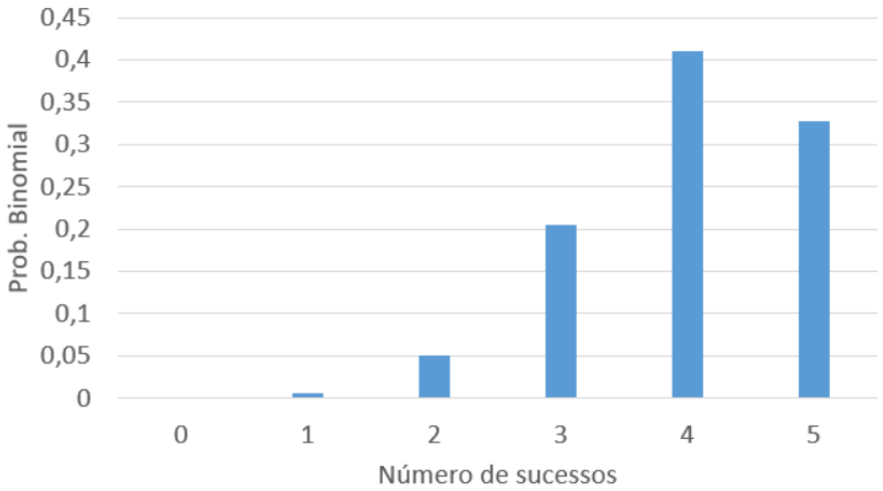
- Na prática, esta solução é simples.
- Contamos o número de sucessos no banco de dados e dividimos pelo total.
- Assumimos no nosso exemplo como P(1) = 0,8

Resolução 2

- Se eu tenho uma base de 5 pessoas para serem cobradas. Qual a chance de cobrar entre 2 a 4 pessoas com sucesso?

Resolução 2

- Vamos primeiro olhar o gráfico:



Qual a área abaixo desta curva? (esta pergunta será base para nossos estudos futuros sobre integrais)

Por enquanto, somente assuma que a base será igual a 1.

A altura será o valor da distribuição binomial.

Resolução 2

- Olhando a tabela vemos o seguinte:

Número de sucessos	Combinações possíveis	Probabilidade Binomial
0	1	$1 \cdot 0{,}2^5 = 0{,}00032$
1	5	$5 \cdot 0{,}8^1 \cdot 0{,}2^4 = 0{,}0064$
2	10	$10 \cdot 0{,}8^2 \cdot 0{,}2^3 = 0{,}0512$
3	10	$10 \cdot 0{,}8^3 \cdot 0{,}2^2 = 0{,}2048$
4	5	$5 \cdot 0{,}8^4 \cdot 0{,}2^1 = 0{,}4096$
5	1	$1 \cdot 0{,}8^5 = 0{,}3277$

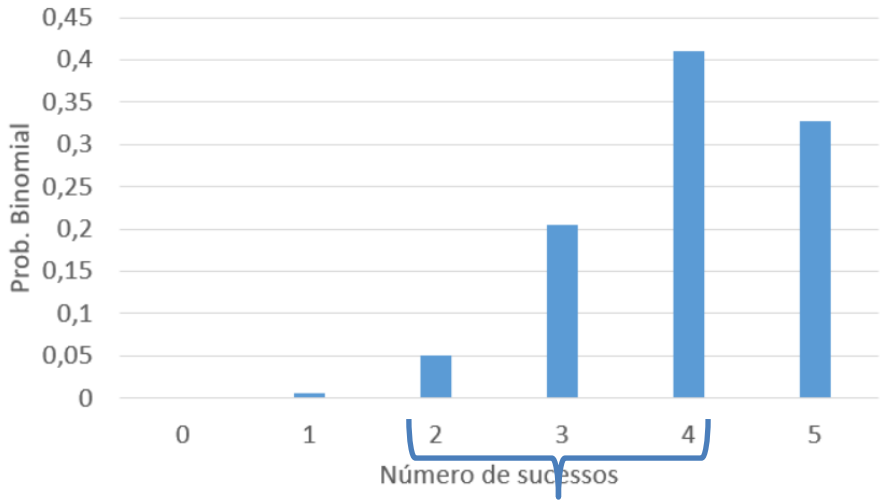
A soma de todos estes valores será igual a 1.

Ou seja, a área total abaixo da curva será igual a 1.

Mas, queremos obter a chance entre 2 a 4 sucessos.

Resolução 2

- Voltando para o gráfico:



Não queremos a área inteira, queremos somente esta área.

Esta área indica a probabilidade entre 2 a 4 cobranças com sucesso.

Resolução 2

- Olhando a tabela vemos o seguinte:

Número de sucessos	Combinações possíveis	Probabilidade Binomial
0	1	$1 \cdot 0{,}2^5 = 0{,}00032$
1	5	$5 \cdot 0{,}8^1 \cdot 0{,}2^4 = 0{,}0064$
2	10	$10 \cdot 0{,}8^2 \cdot 0{,}2^3 = 0{,}0512$
3	10	$10 \cdot 0{,}8^3 \cdot 0{,}2^2 = 0{,}2048$
4	5	$5 \cdot 0{,}8^4 \cdot 0{,}2^1 = 0{,}4096$
5	1	$1 \cdot 0{,}8^5 = 0{,}3277$

Esta área será 0,6656

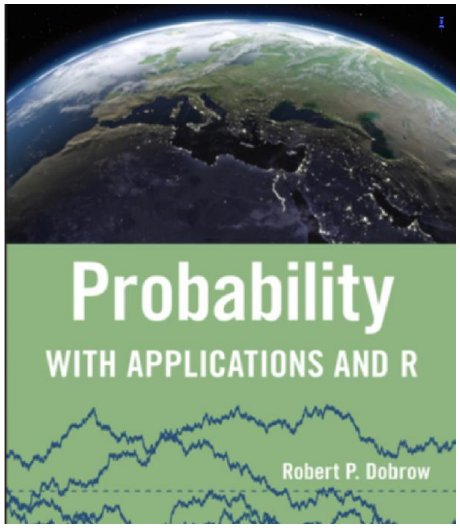
Deste modo, a chance de obter sucesso entre 2 a 4 cobranças é de 66,56%.

Exercícios

- Assistir o vídeo abaixo:
- https://www.youtube.com/watch?v=6uOLbhYeXrk
- Para qual distribuição de probabilidade os exercícios foram resolvidos?

Onde estudar mais!!

- Leitura



- Outras distribuições:
  https://support.minitab.com/pt-br/minitab/18/help-and-how-to/probability-distributions-and-random-data/how-to/probability-distributions/perform-the-analysis/select-the-distribution-and-enter-the-parameters/

- Vídeos

- Variável aleatória:
  https://www.khanacademy.org/math/statistics-probability/random-variables-stats-library/random-variables-discrete/v/random-variables

- Distribuições de probabilidade:
  https://www.khanacademy.org/math/ap-statistics/random-variables-ap/discrete-random-variables/v/discrete-probability-distribution

THE WAY WE ARE