Faculdade XPe

Capítulo 02

ESTATÍSTICA PARA CIÊNCIA DE DADOS

Arnaldo Satoru Gunzi

2024

Leis de Probabilidade

As distribuições de probabilidades são funções matemáticas que nos ajudam a modelar a incerteza sobre variáveis do mundo real. Em ambiente de negócios, a incerteza está contida em diversos momentos, por exemplo: quanto de vendas ocorrerá, qual melhor dia de promoção etc.
A probabilidade é expressa como uma fração ou número decimal entre 0 e 1. Por exemplo, ao jogar uma moeda para cima e observar o resultado, teremos 0,5 (ou 50%) de chances de ter cara ou ter coroa.

Leis de Probabilidade e Diretrizes para sua Aplicação

Para encontrar a probabilidade de um determinado evento ocorrer, podemos utilizar a probabilidade frequentista. Sendo A um evento aleatório qualquer, podemos encontrar a probabilidade de A utilizando a probabilidade frequentista da seguinte forma:

$$P(A) = \frac{Número\ de\ Vezes\ que\ o\ evento\ A\ ocorreu}{Número\ total\ de\ observações}$$

De tal forma que:

$$0 \leq P(A) \leq 1$$

Utilizando um exemplo hipotético, vamos supor que desejamos saber a probabilidade de um cliente realizar uma compra ao entrar em nossa loja. Ao fazer um levantamento dos dados, observamos que 1000 clientes entraram em nossa loja, e desses, 500 compraram. Logo, a probabilidade de um cliente comprar fica:

$$P(Comprar) = \frac{Número\ total\ de\ clientes\ que\ entraram\ e\ compraram}{Número\ total\ de\ clientes\ que\ entraram}$$

$$P(Comprar) = \frac{500}{1000} = 0,5\ (50\%)$$

Ao trabalharmos com probabilidades, é fundamental definirmos o evento de interesse. Nesse caso, o evento de interesse é o cliente comprar. A notação para o evento de interesse (sucesso) ocorrer é o número 1, e para o evento de não interesse (fracasso) é o 0. Em nosso exemplo, temos que:

$$P(1)=P(Comprar)=0,5$$

Portanto, a probabilidade de o evento não ocorrer é o espaço complementar:

$$P(0)=P(Não\ Comprar)=1-P(Comprar)=1-0,5=0,5$$

E se a probabilidade de comprar fosse 0,6? A probabilidade de não comprar seria:

$$P(Não\ Comprar)=1-0,6=0,4$$

Como a probabilidade não passa de 1 (100%), a seguinte propriedade deve ser sempre respeitada:

$$P(Sucesso)+P(Fracasso)=P(1)+P(0)=1$$

Ou seja, a probabilidade de o evento de interesse ocorrer mais a probabilidade de não ocorrer deve fechar em 1 (100%).
Duas regras fundamentais no estudo da teoria das probabilidades são as Regras Aditivas e as Regras Multiplicativas.
Para que não fique abstrato, vamos propor um contexto e, posteriormente, utilizá-lo como exemplo para entender o que são as **Regras Aditivas** e as **Regras Multiplicativas**.
Suponha que você tenha um restaurante em que uma promoção foi lançada. Como cortesia, para cada cliente, será sorteada, de forma aleatória, uma sobremesa. Existem disponíveis dez sobremesas, sendo que:

- 1 sobremesa possui cobertura de menta.
- 2 sobremesas possuem cobertura de chocolate.
- 3 sobremesas possuem cobertura de morango.
- 1 sobremesa possui cobertura de chocolate e cobertura de morango.
- 3 sobremesas possuem cobertura de baunilha.

Qual a probabilidade do cliente receber uma sobremesa com cobertura de menta **ou** uma sobremesa com cobertura de chocolate?
O operador **ou** nos informa que pelo menos um dos eventos deve ocorrer. Podemos observar que o evento "receber sobremesa com cobertura de menta" e o evento "receber sobremesa com cobertura de chocolate" não podem ocorrer ao mesmo tempo, pois nenhuma sobremesa tem as duas coberturas, portanto, eles são chamados de eventos **mutuamente exclusivos**.
Para aplicar a regra aditiva, devemos calcular a probabilidade de cada evento ocorrer.
Temos 10 tipos de sobremesas, em que:

- 1 possui cobertura de menta, portanto a probabilidade do cliente receber aleatoriamente uma sobremesa com cobertura de menta é 1/10 (ou 10%).
- 3 sobremesas possuem cobertura de chocolate, portanto, a probabilidade de o cliente receber aleatoriamente uma sobremesa com cobertura de chocolate é de 3/10 (ou 30%).

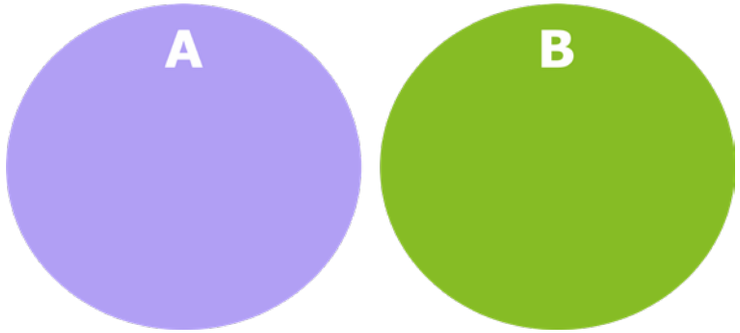
De acordo com a regra aditiva, para obter a probabilidade de um **ou** outro evento aleatório ocorrer, devemos somar suas respectivas probabilidades. Ou seja:

$$P(\text{Menta}) \text{ ou } P(\text{Chocolate}) = P(\text{Menta}) + P(\text{Chocolate})$$

Essa regra significa a união entre os dois eventos:

$$P(\text{Menta}) \cup P(\text{Chocolate})$$

Figura 10 - Eventos mutuamente exclusivos



Como os dois eventos não podem ocorrer ao mesmo tempo, o conjunto intersecção é igual ao conjunto vazio.

$$P(\text{Menta} \cap \text{Chocolate}) = \emptyset$$

Sendo assim, a probabilidade do cliente receber uma sobremesa com cobertura de menta ou uma sobremesa com cobertura de chocolate é:

$$P(\text{Menta}) \text{ ou } P(\text{Chocolate}) = P(\text{Menta}) + P(\text{Chocolate}) = 1/10 + 3/10 = 4/10 \text{ (ou } 40\%)$$

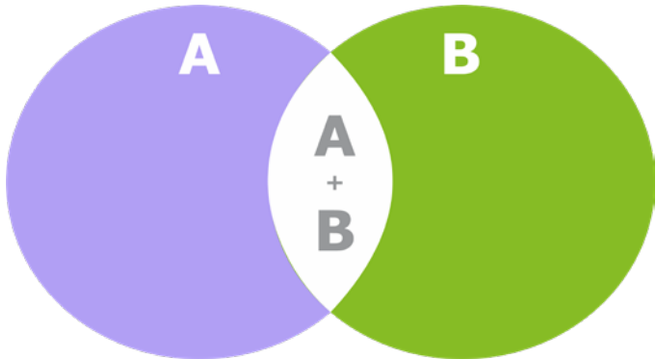
E se desejarmos saber a probabilidade de ocorrência de um **ou** outro evento que não sejam mutuamente exclusivos (ou seja, esses eventos podem ocorrer ao mesmo tempo)? Por exemplo:

Sendo A = rei, B = paus. Sabemos que em um baralho de 52 cartas existem 4 reis, logo p(A) = 4/52. Como também existem 13 cartas de paus, logo P(B)=13/52. Contudo, existe o rei de paus, que é contabilizado tanto em A como em B.
Para o cálculo de P(A ou B), devemos desconsiderar a probabilidade de A e B ocorrerem juntos P(A e B), cujo valor é 1/52, p só há um rei de paus dentre as 52 cartas do baralho.

$$P(\text{A ou B}) = P(\text{A}) + P(\text{B}) - P(\text{A e B}) = 4/52 + 13/52 - 1/52 = 16/52 \text{ (ou } 30{,}76\%)$$

Neste último exemplo, o conjunto intersecção foi removido, pois ele foi contado nos dois conjuntos A e B.

Figura 11 - Removendo a intersecção da união dos dois eventos



Uma vez que compreendemos a **Regra Aditiva** para obter a probabilidade de um **ou** outro evento aleatório ocorrer, vamos aprender como utilizar a **Regra Multiplicativa** para obter a probabilidade de um **e** outro evento aleatório ocorrer.
Supondo que além de sortear aleatoriamente uma das sobremesas para o cliente, também será sorteado um café.
Temos 2 tipos de cafés:

- Expresso
- Cappuccino

Qual a probabilidade do cliente receber uma sobremesa com cobertura de menta e em seguida um café expresso? Queremos obter a probabilidade de um evento ocorrer após um primeiro ter ocorrido, mas a probabilidade de um não influencia na probabilidade do outro (são independentes). Portanto, o cálculo fica:

$$P(\text{Menta}) \text{ e } P(\text{Expresso}) = P(\text{Menta}) * P(\text{Expresso})$$

Como temos 2 tipos de cafés, a probabilidade do expresso ser sorteado é de 1/2 (ou 50%). A probabilidade de uma sobremesa com cobertura de menta ser sorteada já sabemos que é de 1/10. Substituindo na fórmula fica:

$$P(Menta) \text{ e } P(Expresso) = \frac{1}{10} * \frac{1}{2} = \frac{1}{20} (ou\ 5\%)$$

E se os dois eventos foram dependentes, ou seja, dado que um evento ocorreu, a probabilidade do outro se modifica? Vamos supor que o cliente irá receber aleatoriamente duas sobremesas.

Qual a probabilidade de um cliente receber uma sobremesa com cobertura de menta e a seguir a próxima sobremesa sorteada ser de cobertura de baunilha?

Sabemos que 10 tipos de sobremesas podem ser selecionados aleatoriamente, e sabemos que 1 tem cobertura de menta, então a probabilidade de uma sobremesa de cobertura de menta é de 1/10.
Sabemos que existem 3 sobremesas que têm cobertura de baunilha, entretanto temos apenas 9 sobremesas restantes, pois uma já foi sorteada, que foi a que tem cobertura de menta.
Portanto, a probabilidade de sair aleatoriamente uma sobremesa com cobertura de baunilha dado que uma sobremesa com cobertura de menta já saiu, fica 3/9 (ou 33,33%). Observe que o valor do denominador não é mais 10, e sim, 9.
Ou seja, a probabilidade de uma sobremesa com cobertura de baunilha ser sorteada após uma sobremesa com cobertura de menta ser sorteada fica:

$$P(Menta) * P(Baunilha \mid Menta) = \frac{1}{10} * \frac{3}{9} = \frac{3}{90} (ou\ 3\%)$$

Exercícios adicionais de probabilidades

Seguem alguns exercícios resolvidos, para fixação de conceitos.

Exercício: Uma moeda honesta é lançada 3 vezes. Qual a probabilidade de obtermos 2 caras e 1 coroa, em qualquer ordem?

Resposta: Seja C = cara e K = coroa. Espaço amostral: {CCC, CCK, CKC, KCC, KKC, KCK, CKK, KKK} (8 possibilidades)

Eventos favoráveis: {CCK, CKC, KCC} (3 possibilidades)

Probabilidade: 3/8

Exercício: Dois dados honestos são lançados. Qual a probabilidade da soma dos resultados ser 7?

Resposta: Espaço amostral: 6 x 6 = 36 possibilidades

Eventos favoráveis: {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)} (6 possibilidades)

Probabilidade: 6/36 = 1/6

Exercício: Uma urna contém 5 bolas vermelhas, 3 bolas azuis e 2 bolas amarelas. Uma bola é retirada ao acaso. Qual a probabilidade de a bola não ser vermelha?

Resposta: Eventos favoráveis (não vermelhas): 3 + 2 = 5

Probabilidade: 5/10 = 1/2

Eventos Independentes: Em uma escola, a probabilidade de um aluno gostar de matemática é de 0,6 e a probabilidade de gostar de futebol é de 0,7. Se esses eventos são independentes, qual a probabilidade de um aluno gostar de ambos?

Resposta: Como são eventos independentes e queremos a intersecção de ambos os estados, temos 0,6 * 0,7 = 0,42.

Variáveis Aleatórias Discretas e Contínuas

Antes de entrarmos nas distribuições de probabilidades, temos que entender os dois tipos mais importantes de variáveis aleatórias (ou v.a.).

Uma variável aleatória é classificada conforme a natureza do conjunto de valores que ela pode assumir. Os dois tipos de variáveis aleatórias mais importantes são as **variáveis aleatórias discretas** e as **variáveis aleatórias contínuas.**

- **Variável Aleatória Discreta:**

Usualmente, os valores de uma v.a. discreta são oriundos de um processo de contagem. Assumem valores inteiros. Exemplos são quantidade de ligações por dia de call center, quantidade de clientes que entram em uma loja etc.

- **Variável Aleatória Contínua:**

Os valores de uma v.a. contínua assumem um número infinito incontável de valores. Assumem números reais. Geralmente, são oriundos de algum processo de medição. Exemplos: Tempo (em minutos) das ligações que um call center recebe, valor (em reais) que os clientes compram em uma loja etc.

Nos próximos tópicos, serão descritas as principais variáveis aleatórias discretas e contínuas.

Distribuições Discretas

A seguir, algumas das principais distribuições de probabilidades discretas.

- **Experimento de Bernoulli:**

É simplesmente a realização de uma tentativa de um experimento aleatório. Por exemplo, jogar a moeda para cima uma vez e observar se saiu cara ou coroa.

$$\text{p(sucesso)} = \text{p(1)} = \text{p}$$
$$\text{p(fracasso)} = \text{p(0)} = 1\text{-p}$$

Onde **p** é a probabilidade de o sucesso ocorrer.

Vamos definir como sucesso para nosso exemplo sair coroa na face de cima da moeda. Sabemos que a moeda tem duas faces, uma cara e uma coroa. Então, as chances de uma coroa ocorrer é de p, e as chances do sucesso não ocorrer, ou seja, do fracasso, é 1-p. É o pesquisador que define qual evento será considerado o sucesso para determinado estudo.

- **Distribuição Binomial:**

É o número de x sucessos em n tentativas. É a repetição de n experimentos de Bernoulli. A função de probabilidade da distribuição Binomial é:

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

Onde **n** é o número de tentativas, **x** é o número de sucessos, **p** é a probabilidade de sucesso.

Exemplo: Definindo como sucesso o cliente comprar, e supondo que a probabilidade de sucesso é 50% **(p)**. Ao passar 10 **(n)** clientes em nossa loja, qual a probabilidade de realizarmos 2 **(x)** vendas? Vamos substituir os valores na função de probabilidade da distribuição Binomial:

$$f(x) = \binom{10}{2} 0{,}5^2 (1-0{,}5)^8$$
$$f(x) = 0{,}0439 \text{ (ou } 4{,}39\%)$$

- **Distribuição Geométrica:**

É repetir um experimento de Bernoulli x vezes até que o primeiro sucesso ocorra. Ou seja, é o número de fracassos até o primeiro sucesso.

A função de probabilidade da distribuição Geométrica é:

$$f(x) = (1-p)^{x-1}.p$$

Onde **x** é o número de tentativas, **p** é a probabilidade de sucesso.

Exemplo: Definindo como sucesso o cliente comprar, e supondo que a probabilidade de sucesso é 50%(p). Qual a probabilidade da primeira venda ocorrer quando o quinto (x) cliente entrar na loja?

$$f(x) = (1-0{,}5)^{5-1}.0{,}5$$
$$f(x) = 0.03125 \text{ (ou } 3{,}12\%)$$

- **Distribuição Binomial Negativa:**

É o número de x experimentos de Bernoulli até que uma quantidade r de sucessos ocorra. Pode ser vista como uma generalização da distribuição geométrica.

A função de probabilidade da distribuição Binomial Negativa é:

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

Onde **r** é a quantidade de sucessos, **x** é o número de tentativas, **p** é a probabilidade de sucesso.

Exemplo: Definindo como sucesso o cliente comprar, e supondo que a probabilidade de sucesso é 50% **(p)**. Qual a probabilidade de ter que entrar 8 **(x)** clientes até que a segunda **(r)** venda ocorra?

$$f(x) = \binom{8-1}{2-1} (1-0{,}5)^{8-2}.0{,}5^2$$
$$f(x) = 0{,}02734 \text{ (ou } 2{,}73\%)$$

- **Distribuição de Poisson:**

Expressa a probabilidade de um evento ou uma série de eventos ocorrerem em um determinado período de tempo ou espaço.
A função de probabilidade de distribuição de Poisson é:

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Onde **e**=2,71 é o número de Euler, **x!** é o fatorial de número de vezes que o evento ocorre, **λ** é o número de ocorrências de um evento aleatório em um determinado intervalo de tempo ou espaço.
Exemplo: Uma loja recebe em média, 6 (**λ**) clientes por minuto. Qual a probabilidade de que 5(**x**) clientes entrem em um minuto?

$$f(x) = \frac{e^{-6}6^{5}}{5!} = 0{,}1606 \text{ (ou 16,06%)}$$

Distribuições Contínuas

Vamos mostrar algumas distribuições de probabilidades contínuas.

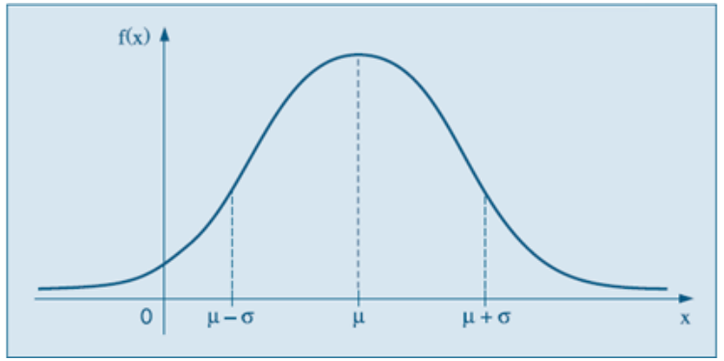
- **Distribuição Normal (Gaussiana)**

É uma das distribuições mais importantes da Estatística. A curva Normal ou Gaussiana descreve de forma muito adequada o comportamento de uma variável aleatória que se distribui de forma simétrica em relação a um valor central. Os dois parâmetros que a caracterizam são a média μ (que especifica o valor central) e a variância σ^2 (que define sua variabilidade em torno da média).
Função de densidade de uma distribuição normal:

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

Onde, **x** é o valor da variável aleatória, **μ** é a média, σ é o desvio padrão, π = 3,14, e = 2,71.

Figura 12 - Distribuição Normal ou Gaussiana



Se uma v.a. chamada de X segue uma distribuição normal com média μ e desvio padrão σ, podemos representar pela notação: **X ~ N(μ,** σ**)**.
Na prática, vamos utilizar pacotes computacionais para realizar esses cálculos. Acompanhe os códigos indicados no final desta seção.

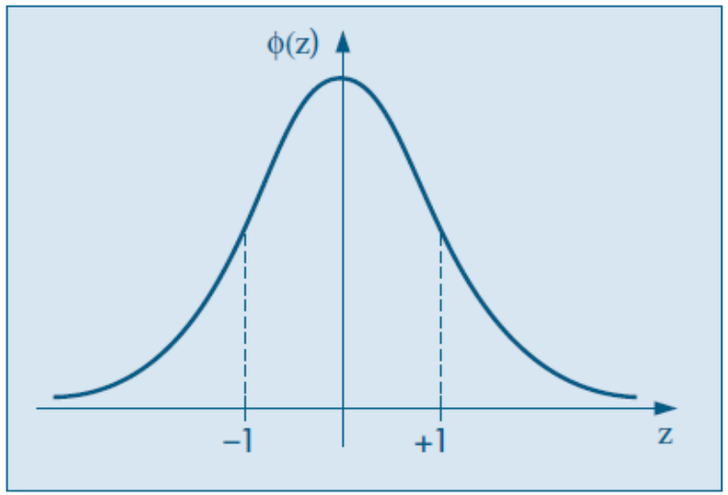
- **Distribuição Normal Padrão** (distribuição z)**:**

É um caso especial da distribuição normal. Se uma variável aleatória segue uma distribuição normal, uma transformação (chamada de padronização) é aplicada de modo que essa variável tenha média zero e desvio padrão unitário. A equação a seguir demonstra como transformar uma variável aleatória normal em uma variável Z.

$$z = \frac{(x_i - \mu)}{\sigma}$$

Onde **x** é o i-ésimo valor da v.a., **μ** é a média da v.a. e σ é o desvio padrão da v.a.

Figura 13 - Distribuição Normal Padrão (distribuição Z)



Se uma v.a. chamada de X segue uma distribuição normal padrão com média zero e desvio padrão unitário, utilizamos a notação: **X ~ Z(μ=0, σ=1)**
Vamos aproveitar o contexto anterior para aplicar a distribuição Z e a sua tabela de probabilidades.

- **Outras distribuições**

Há diversas outras distribuições de probabilidades notáveis, que não serão cobertas neste curso. Ao aluno interessado, pesquisar por Distribuição F de Fisher, T de Student e Qui-Quadrado, dentre outros.

Pílulas de conhecimento

O Geogebra é uma excelente ferramenta computacional para visualização de distribuições de probabilidades. Vide: https://www.geogebra.org/m/dPqN3u5p

Estatística Computacional – Probabilidades com o Python

Vide link a seguir para acompanhar o código para análise de probabilidades com Python.
Link do Github: https://github.com/asgunzi/Estatistica_Analise_Dados

Faculdade XPe

xpeducacao.com.br