Universidade de Brasília


EXPERIMENTAÇÃO: 3 - DISTRIBUIÇÕES


RESUMO

Descreve os principais tipos de distribuição



User

Concepta McManus

Sonia Nair Báo

Daniel Pimentel

Felipe Pimentel

Andrea Queiroz Maranhão

José Jivago Rolo

## Sumário

## Os principais tipos de distribuição estatísticas

Todo livro de estatística fornece uma lista de distribuições estatísticas, com suas propriedades, mas navegar por essas opções pode ser frustrante para qualquer pessoa sem conhecimento estatístico, por dois motivos. Primeiro, as escolhas parecem infinitas, com dezenas de distribuições competindo por sua atenção, com pouca ou nenhuma base intuitiva para diferenciá-las. Em segundo lugar, as descrições tendem a ser abstratas e enfatizam propriedades estatísticas como os momentos, funções características e distribuições cumulativas. Neste apêndice, vamos nos concentrar nos aspectos das distribuições que são mais úteis ao analisar dados brutos e tentar ajustar a distribuição correta a esses dados.

De uma perspectiva prática, podemos pensar em uma distribuição como uma função que descreve a relação entre observações em um espaço amostral.

Por exemplo, podemos estar interessados na idade dos humanos, com idades individuais representando observações no domínio e idades de 0 a 125 a extensão do espaço amostral. A distribuição é uma função matemática que descreve a relação de observações de diferentes alturas.

*Funções de densidade*

As distribuições são frequentemente descritas em termos de suas funções de densidade ou densidade.

As funções de densidade são funções que descrevem como a proporção de dados ou a probabilidade da proporção de observações mudam ao longo do intervalo da distribuição.

Dois tipos de funções de densidade são funções de densidade de probabilidade e funções de densidade cumulativa.

- Função densidade de probabilidade: calcula a probabilidade de observar um determinado valor.

- Função de densidade cumulativa: calcula a probabilidade de uma observação igual ou menor que um valor.

Uma função de densidade de probabilidade, ou PDF, pode ser usada para calcular a probabilidade de uma determinada observação em uma distribuição. Também pode ser usado para resumir a probabilidade de observações em todo o espaço amostral da distribuição. Os gráficos do PDF mostram a forma familiar de uma distribuição, como a curva de sino para a distribuição gaussiana.

As distribuições são frequentemente definidas em termos de suas funções de densidade de probabilidade com seus parâmetros associados.

Uma função de densidade cumulativa, ou CDF, é uma maneira diferente de pensar sobre a probabilidade de valores observados. Em vez de calcular a probabilidade de uma determinada observação como no PDF, o CDF calcula a probabilidade cumulativa para a observação e todas as observações anteriores no espaço amostral. Ele permite que você entenda e comente rapidamente quanto da distribuição está antes e depois de um determinado valor. Um CDF é frequentemente plotado como uma curva de 0 a 1 para a distribuição.

Ambos PDFs e CDFs são funções contínuas. O equivalente de um PDF para uma distribuição discreta é chamado de função de massa de probabilidade, ou PMF.

**Estatísticas para Descrever a Distribuição**

Usando Distribuições Padrão como Distribuições de Referência

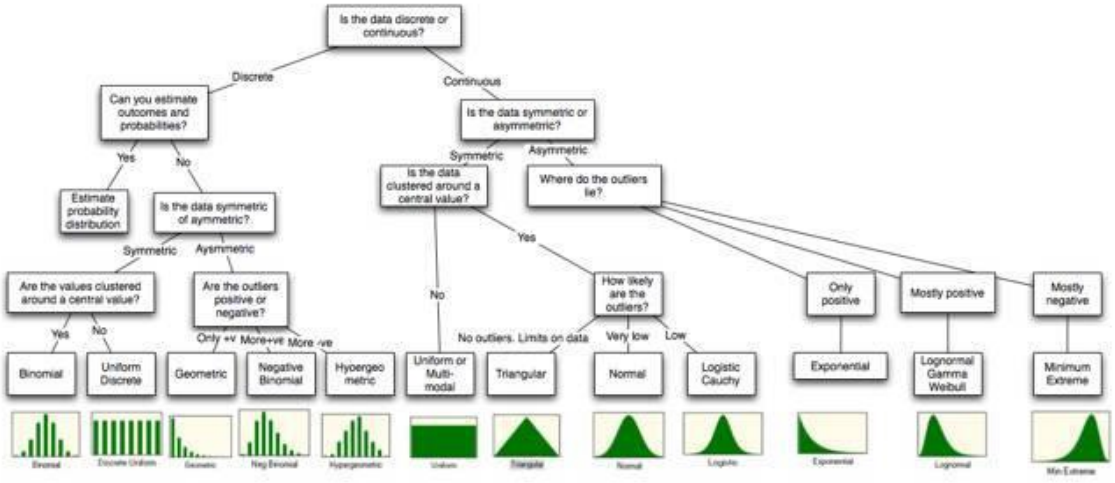
As distribuições padrão são frequentemente usadas como distribuições de referência em testes estatísticos.

Isso significa que os dados da amostra são comparados com eles para ver a probabilidade de que os dados tenham ocorrido aleatoriamente.

As características das distribuições padrão as tornam muito adequadas para serem distribuições de referência, especialmente as características bem conhecidas, e o fato de serem boas aproximações de dados do mundo real.

No entanto, existem outras fontes de distribuições de referência.

- As distribuições de bootstrap são criadas assumindo que os dados de amostra são os únicos dados disponíveis e desenhando amostras repetidas (menores) desses dados. Eles só podem realmente ser usados quando você tem acesso a um computador e não são ideais. Portanto, eles devem ser usados apenas quando não houver alternativa.
- As distribuições permutacionais são criadas encontrando todas as permutações possíveis de dados classificados. Eles, portanto, pegam todos os resultados possíveis e veem quão prováveis eles são. Eles não assumem qualquer distribuição teórica subjacente. Testes que usam essas distribuições são conhecidos como testes 'não paramétricos', para distingui-los dos testes 'paramétricos' que usam distribuições padrão com parâmetros conhecidos.
- Os dados de arquivo também podem ser usados para criar uma distribuição de referência. Isso pode ser apropriado onde há muitos dados anteriores que podem ser usados.




## Por que as distribuições estatísticas são importantes

A principal razão pela qual você precisa entender sobre distribuições estatísticas é seu uso em testes estatísticos. Você pode usá-los para comparar seus dados, para ajudá-lo a entender a probabilidade de você ter identificado um relacionamento ou recurso real de seus dados.

A questão de qual distribuição melhor se ajusta aos dados não pode ser respondida sem observar se os dados são discretos ou contínuos, simétricos ou assimétricos e onde estão os outliers.

## Testes de ajuste[1]

O teste mais simples para ajuste distribucional é visual com uma comparação do histograma dos dados reais com a distribuição ajustada.

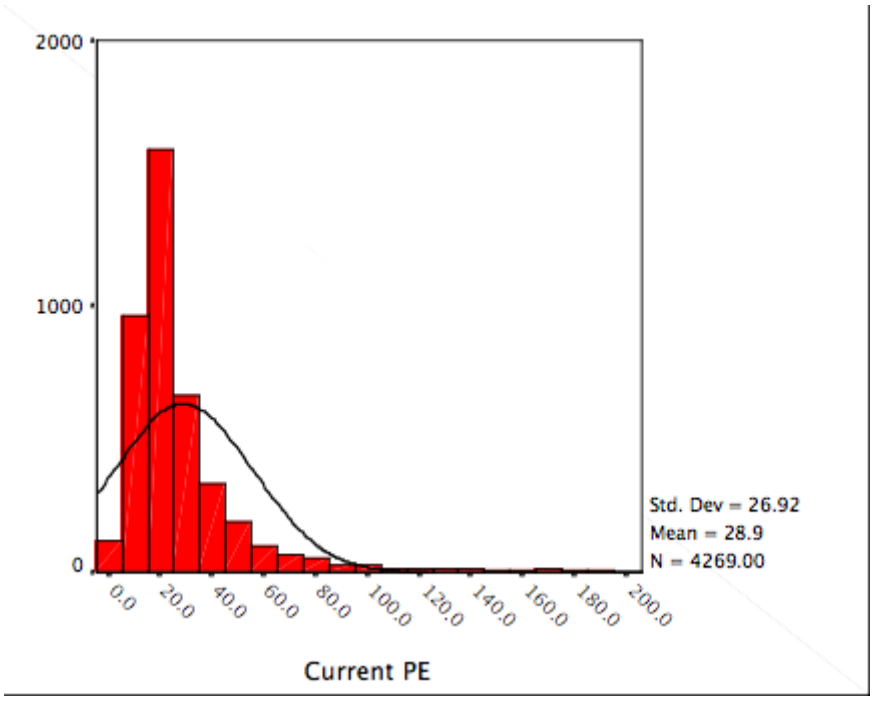



Figura. Distribuição dos atuais índices preço/lucro para ações dos EUA, com uma distribuição normal sobreposta.

As distribuições são tão claramente divergentes que a suposição de distribuição normal não se sustenta.

Um teste um pouco mais sofisticado é calcular os momentos da distribuição real dos dados – a média, o desvio padrão, assimetria e curtose – e examiná-los para ajuste à distribuição escolhida. Com os dados de preço-lucro acima, por exemplo, os momentos da distribuição e as principais estatísticas são resumidos na tabela.

[1] https://education.ti.com/en/building-concepts/activities/statistics/sequence1/analyzing-distributions

| | *Current PE* | *Normal Distribution* |
|---|---|---|
| Mean | 28.947 | |
| Median | 20.952 | Median = Mean |
| Standard deviation | 26.924 | |
| Skewness | 3.106 | 0 |
| Kurtosis | 11.936 | 0 |

Como a distribuição normal não tem assimetria e zero curtose, podemos facilmente rejeitar a hipótese de que as relações preço-lucro são normalmente distribuídas.

Os testes típicos de bondade de ajuste comparam a função de distribuição real dos dados com a função de distribuição cumulativa da distribuição que está sendo usada para caracterizar os dados, para aceitar a hipótese de que a distribuição escolhida se ajusta aos dados ou para rejeitá-la. Não surpreendentemente, dado seu uso constante, há mais testes de normalidade do que para qualquer outra distribuição. O teste de Kolmogorov-Smirnov é um dos mais antigos testes de ajuste para distribuições, datado de 1967. Versões melhoradas dos testes incluem os testes de Shapiro-Wilk e Anderson-Darling.

O teste de *Kolgomorov-Smirnov* pode ser usado para ver se os dados se ajustam a uma distribuição normal, lognormal, Weibull, exponencial ou logística.

*Estatística de Anderson-Darling (AD):* Existem diferentes testes de distribuição. O teste que usarei para nossos dados é o teste de Anderson-Darling. A estatística de Anderson-Darling é a estatística de teste. É como o valor t para testes t ou o valor F para testes F. Normalmente, você não interpreta essa estatística diretamente, mas o software a usa para calcular o valor p para o teste.

*Valor-P:* Testes de distribuição que possuem valores-p altos são candidatos adequados para a distribuição de seus dados. Infelizmente, não é possível calcular valores de p para algumas distribuições com três parâmetros.

*LRT P:* Se você estiver considerando uma distribuição de três parâmetros, avalie o LRT P (Valor de P para Likelihood Ratio Test) para determinar se o terceiro parâmetro melhora significativamente o ajuste em comparação com a distribuição de dois

parâmetros associada. Um valor LRT P inferior ao seu nível de significância indica uma melhoria significativa em relação à distribuição de dois parâmetros. Se você vir um valor mais alto, considere ficar com a distribuição de dois parâmetros[2].

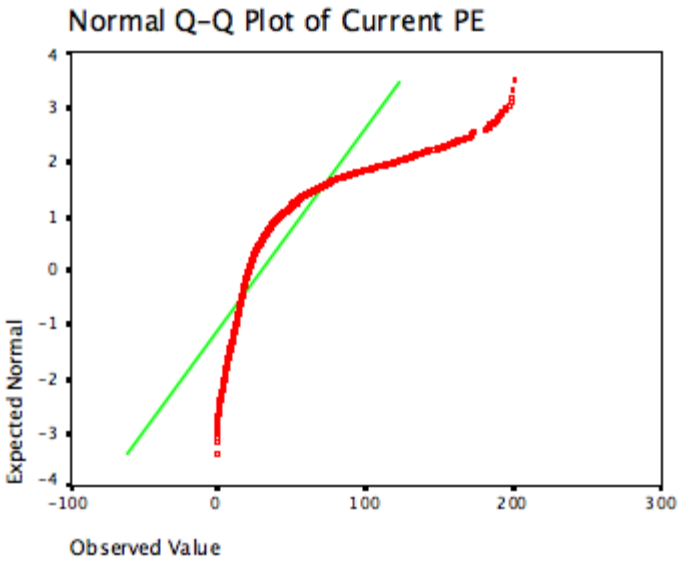
*Akaike information criterion (AIC*): AIC compara a "qualidade" relativa de um modelo (distribuição) versus os outros modelos. Você pode usar o AIC para selecionar a distribuição que melhor se ajusta aos dados. A distribuição com o menor valor de AIC é geralmente o modelo preferido. AIC é definido como:

$$AIC = 2k - 2(\text{Log-Probabilidade})$$

onde k é o número de parâmetros. Observe que o valor AIC sozinho para uma única distribuição não nos diz nada. Não é um teste como o valor-p da estatística de Anderson-Darling. O valor AIC compara a qualidade relativa de todas as distribuições. Portanto, se todas as distribuições não se ajustarem bem aos dados, o valor do AIC não permitirá que você saiba disso. Você precisa combinar os valores p para a estatística Anderson-Darling, o LRT e o valor AIC para ajudar a determinar quais dados se ajustam melhor à distribuição.

Existem testes gráficos de normalidade, onde gráficos de probabilidade podem ser usados para avaliar a hipótese de que os dados são extraídos de uma distribuição normal.

[2] https://www.real-statistics.com/distribution-fitting/




Dado que a distribuição normal é uma das mais fáceis de trabalhar, é útil começar testando os dados quanto à não normalidade para ver se você pode usar a distribuição normal. Caso contrário, você pode estender sua pesquisa para outras distribuições mais complexas.

### *Ajustando a distribuição*[3]

Quando confrontado com dados que precisam ser caracterizados por uma distribuição, é melhor começar com os dados brutos e responder a quatro perguntas básicas sobre os dados que podem ajudar na caracterização. A primeira diz respeito se os dados podem assumir apenas valores discretos ou se os dados são contínuos; se um novo medicamento farmacêutico obtém ou não a aprovação do FDA é um valor discreto, mas as receitas do medicamento representam uma variável contínua. A segunda analisa a simetria dos dados e se há assimetria, em que direção se encontra; em outras palavras, são outliers positivos e negativos igualmente prováveis ou um é mais provável que o outro. A terceira questão é se existem limites superiores ou inferiores nos dados;; existem alguns itens de dados, como receitas, que não podem ser inferiores a zero, enquanto outros, como margens operacionais, não podem exceder um valor (100%). A questão final e correlata diz respeito à probabilidade de se observar

[3] https://www.spcforexcel.com/knowledge/basic-statistics/deciding-which-distribution-fits-your-data-best

valores extremos na distribuição; em alguns dados, os valores extremos ocorrem com pouca frequência, enquanto em outros ocorrem com mais frequência.

### Os dados são discretos ou contínuos?

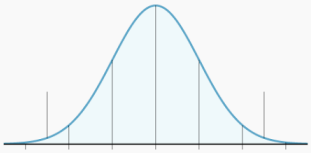
A primeira e mais óbvia categorização de dados deve ser se os dados se restringem a assumir apenas valores discretos ou se são contínuos. Considere as entradas em uma análise de projeto típica em uma empresa. A maioria das estimativas que entram na análise vem de distribuições contínuas; tamanho do mercado, participação de mercado e margens de lucro, por exemplo, são variáveis contínuas. Existem alguns fatores de risco importantes, porém, que podem assumir apenas formas discretas, incluindo ações regulatórias e a ameaça de um ataque terrorista; no primeiro caso, a entidade reguladora pode dispensar uma de duas ou mais decisões previamente especificadas e, no segundo, está sujeito a um ataque terrorista ou não está.

Com dados discretos, toda a distribuição pode ser desenvolvida do zero ou os dados podem ser ajustados a uma distribuição discreta pré-especificada. Com o primeiro, há duas etapas para construir a distribuição. A primeira é identificar os resultados possíveis e a segunda é estimar as probabilidades de cada resultado. Como observamos no texto, podemos nos basear em dados históricos ou experiência, bem como em conhecimentos específicos sobre o investimento que está sendo analisado para chegar à distribuição final. Esse processo é relativamente simples de realizar quando há poucos resultados com uma base bem estabelecida para estimar probabilidades, mas torna-se mais tedioso à medida que o número de resultados aumenta.

# TOP 10

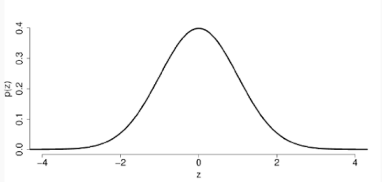
## Types of Distribution in Statistics With Formulas

### Normal Distribution



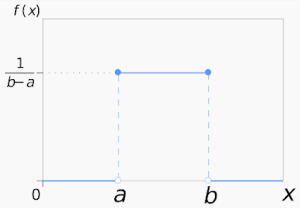
Sometimes, the normal distribution is also called the bell curve. It occurs naturally in several cases; for example, the normal distribution can be seen in tests such as GRE and SAT. Furthermore, there are several groups that follow the normal distribution pattern.

### T- Distribution




It is one of the most important distribution in statistics. It is also known as Student's t- distribution, which is the probability distribution. That is used to estimate the parameters of the population when the given sample size is small.

### Uniform Distribution




The basic form of a continuous distribution is known as uniform distribution. It has the constant probability that forms a rectangular distribution. And it implies that each value has the same length of distribution.

**Characteristics of uniform distribution**

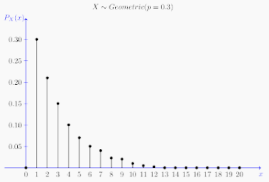
The density function combines to unity.
Each of its input function has equal weightage.
The uniform function's mean is given by:

$$\mu = \frac{(a+b)}{2}$$

The variance of the uniform distribution is given by:

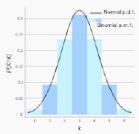
$$V(x) = \frac{(b-a)^2}{12}$$

### Bernoulli distribution



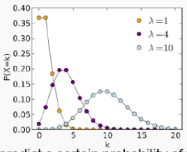
The basic form of a continuous distribution is known as uniform distribution. It has the constant probability that forms a rectangular distribution. And it implies that each value has the same length of distribution.

### Binomial distribution



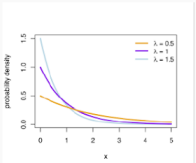
It is a probability distribution that concludes the value that takes one of two independent values under a set of assumptions or parameters. Besides, the binomial distribution's assumptions must have a single result with the same probability of success.

### Poisson distribution




It is a tool that is used to predict a certain probability of the event when you know the value of happening of a certain event.

### Exponential distribution




It is also known as a negative exponential distribution that represents the time between the trails in a Poisson process.

**Some of the formulas of it**

An exponential random variable's expected value is given by:

$$E[X] = \frac{1}{\lambda}$$

An exponential random variable's variance value is given by:

$$\mathrm{Var}[X] = \frac{1}{\lambda^2}$$

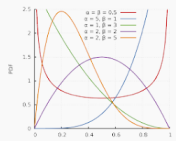
The exponential random variable's moment generating function is given by:

$$M_X(t) = \frac{\lambda}{\lambda - t}$$

An exponential random variable's characteristic function is given by:

$$\varphi_X(t) = \frac{\lambda}{\lambda - it}$$

### Beta distribution



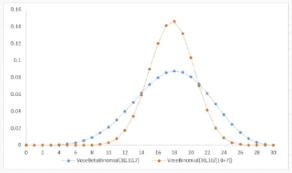
It is the family of continuous probability distributions that set under the interval [0,1], which is expressed by alpha and beta.

**Properties of beta distribution**

The terms to measure the central tendency are:

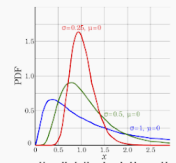
Mean
Harmonic Mean
Mode
Median
Geometric Mean

### Beta-binomial distribution



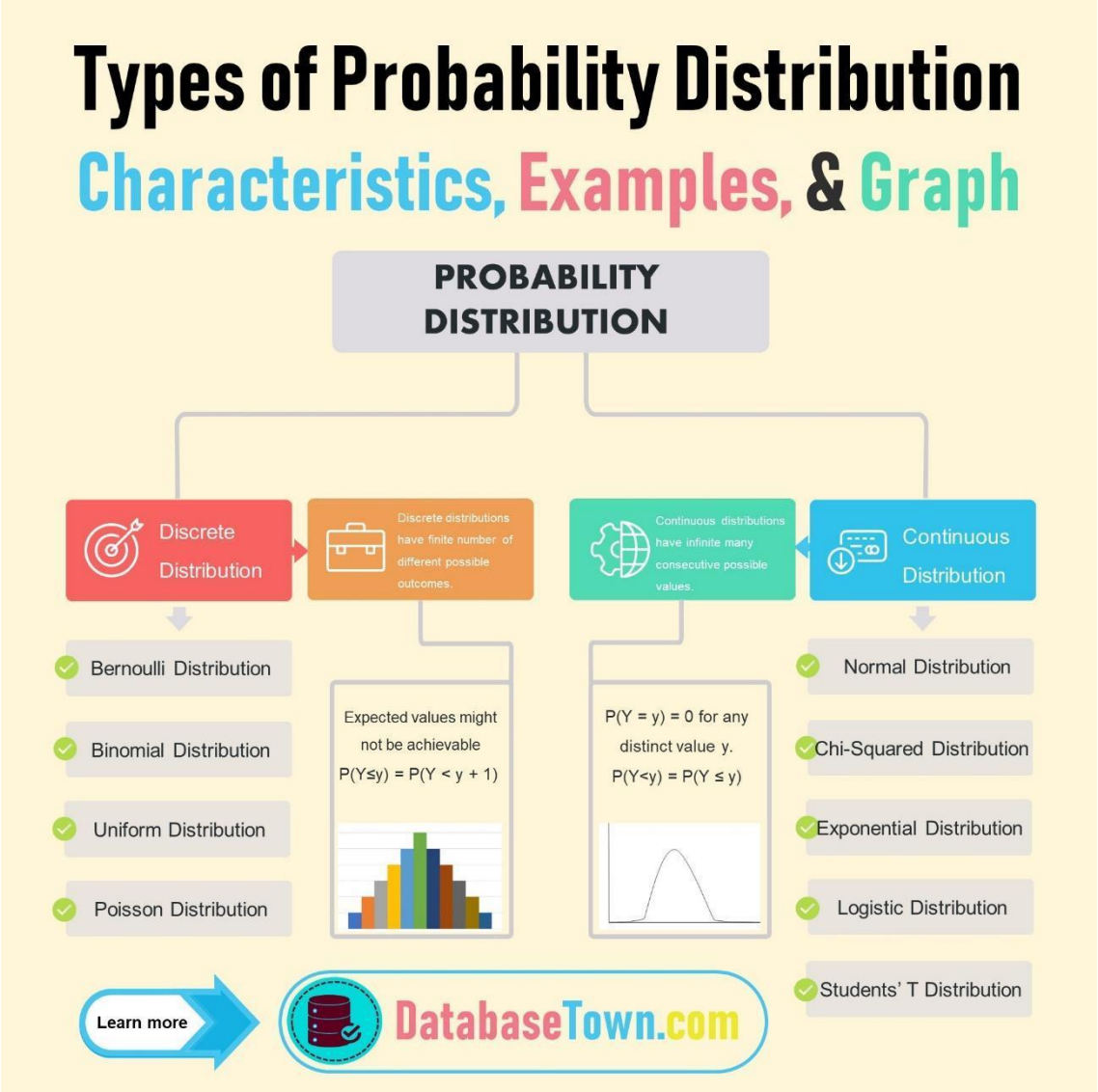
It is the simplest Bayesian model that is widely used in intelligence testing, epidemiology, and marketing. A distribution is said to be beta-binomial.

### Log-normal distribution




f the log to the power is normally distributed, then the variable is taken as lognormally distributed. Or we can say that ln(x) is normally distributed and that the variable x is assumed to have a log-normal distribution.

Existem vários tipos de distribuição de estatísticas, e cada livro os relaciona com suas propriedades. Mas existem vários alunos que ficam frustrados com todos esses tipos; isso se deve a duas razões. O primeiro é que os tipos podem parecer infinitos. Além disso, cada um deles deve ser levado em consideração individualmente. Além disso, o segundo motivo são as descrições que tendem a indicar as propriedades das estatísticas como funções de personagem, momentos e distribuições cumulativas.



**Distribuições contínuas**

Beta

Cauchy

Chi-Square

Dirichlet

Exponential

Extreme value distribution, or the Gumbel

F distribution

Gamma

Logistic

Lognormal

Normal

Pareto

Student's t

Triangular

Weibull

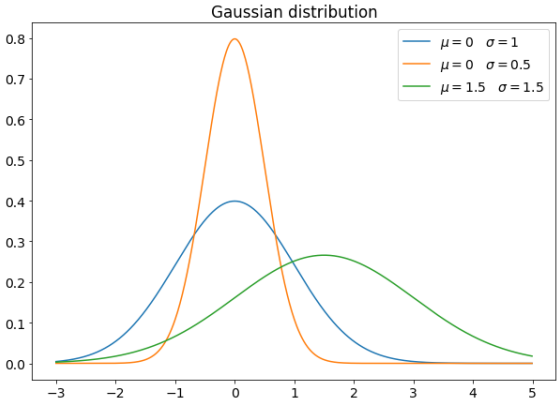
**Distribuição gaussiana**

A rainha de todas as distribuições e talvez a mais conhecida. Se você tiver algumas variáveis aleatórias e independentes com variância finita, a distribuição de probabilidade de sua soma converge para uma distribuição gaussiana.

Esta é a função de densidade de probabilidade:

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

μ e σ são a média e o desvio padrão da distribuição. A média é a posição do pico, enquanto o desvio padrão está relacionado à largura. Quanto maior o desvio padrão, maior será a distribuição (e menor será o pico, para manter a área igual a 1).



**Distribuição normal**

Um caso particular de distribuição gaussiana é a distribuição normal, cuja média é igual a 0 e o desvio padrão é igual a 1.

Cada modelo de regressão usando a abordagem de mínimos quadrados assume que os resíduos são normalmente distribuídos, por isso é importante saber se nossos dados são normalmente distribuídos. O teste F para comparar as variâncias dos resíduos de um modelo em dois conjuntos de dados assume que os resíduos são normalmente distribuídos.

Às vezes, a distribuição normal também é chamada de curva em sino. Ocorre naturalmente em vários casos. Além disso, existem vários grupos que seguem o padrão de distribuição normal. Por causa disso, é amplamente utilizado em estatísticas, negócios e órgãos governamentais como o FDA:

- Erros de medição.
- Pontos em um teste.
- Salários.
- Altura das pessoas.
- Pressão sanguínea.
- Pontuações de QI

*Propriedades de uma distribuição normal*

- A curva permanece simétrica no centro.
- A área sob a curva é 1.
- A média, mediana e moda são sempre iguais.
- Exatamente a metade do valor está à esquerda do centro e a outra à direita.

Por ser uma distribuição contínua, a distribuição normal é mais comumente usada em ciência de dados. Um processo muito comum no nosso dia a dia pertence a esta distribuição - distribuição de renda, relatório médio de funcionários, peso médio de uma população, etc.

A fórmula para distribuição normal;

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

Onde μ = valor médio,

σ = distribuição de probabilidade padrão de probabilidade,

x = variável aleatória

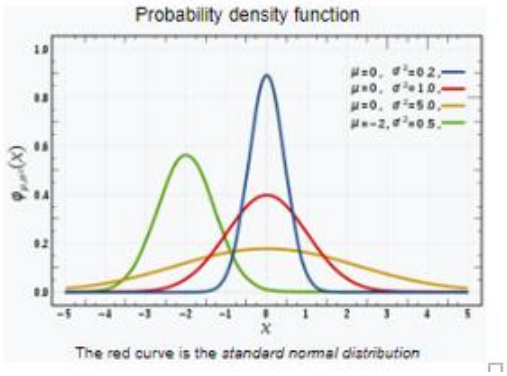
De acordo com a fórmula, a distribuição é dita normal se média (μ) = 0 e desvio padrão (σ) = 1

O gráfico da distribuição normal é mostrado abaixo, o qual é simétrico em relação ao centro (média).

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for all values of } x \text{ and } \mu; \text{ while } \sigma > 0$$

$$Mean = \mu$$

$$Standard\ Deviation = \sigma$$



**Distribuição t de Student**

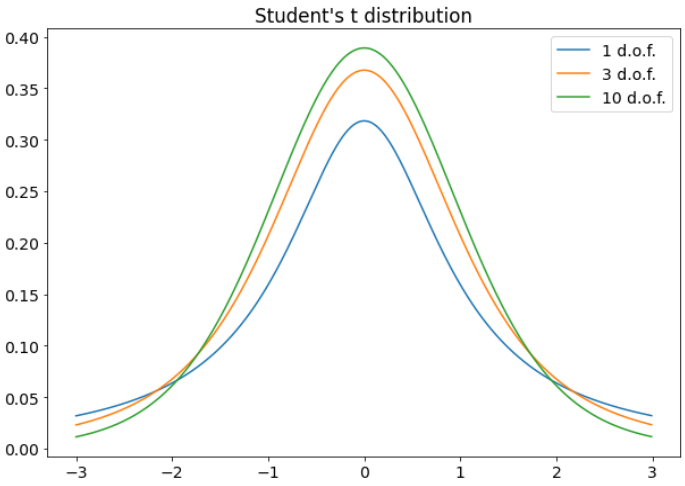
- É uma das distribuições mais importantes em estatísticas. Também é conhecido como distribuição t de Student, que é a distribuição de probabilidade. Isso é usado para estimar os parâmetros da população quando o tamanho da amostra fornecido é pequeno. E o desvio padrão da população é desconhecido. Em estatísticas, a distribuição t é a distribuição mais importante.

*Função de densidade*

$$f(t,n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-(\frac{n+1}{2})}$$

O parâmetro livre n é chamado de "graus de liberdade" (geralmente abreviado em gl). É fácil mostrar que a distribuição t se aproxima de uma distribuição normal para altos valores de n.

É amplamente utilizado para testes de hipóteses e intervalos de confiança construídos para valores médios. O gráfico da distribuição da distribuição t é mostrado abaixo;



*Propriedades da distribuição t*

• Como a distribuição normal, a distribuição do aluno tem formato de sino e simétrica com média zero.

• A faixa de distribuição do aluno de - ∞ a ∞ (infinito).

• A forma da distribuição t muda com o grau de liberdade.

• A variância é sempre mais de um, e pode ser representada quando o grau de liberdade V> = 3 e dado: Var (t) = [v / v -2].

• Não é muito embalado no centro, mas mais alto nas tentativas; portanto, sua forma é platicúrtica.

• A dispersão da distribuição t é muito mais do que a distribuição normal. À medida que o tamanho da amostra 'n' aumenta, é considerada uma distribuição normal. Aqui, o tamanho da amostra fornecido é considerado maior do que n> = 30.

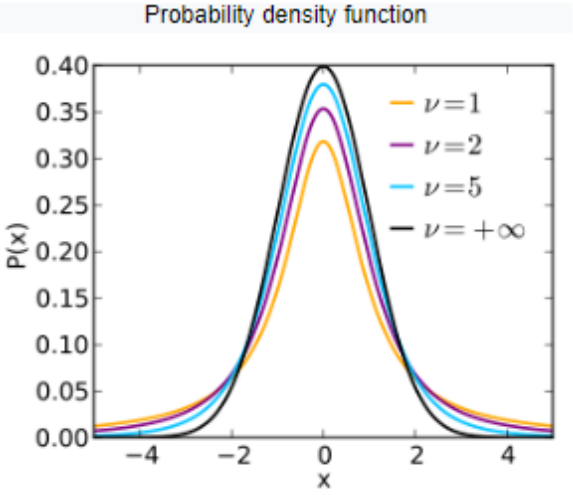
- normalmente usado para testar a significância estatística da diferença entre duas médias amostrais ou para estimar a média de uma população normalmente distribuída, ambos para tamanhos de amostra pequenos. A forma desta distribuição assemelha-se à forma de sino de uma gaussiana leptocúrtica. O único parâmetro é o grau de liberdade r.

$$f(t) = \frac{\Gamma[(r+1)/2]}{\sqrt{r\pi}\,\Gamma[r/2]}(1 + t^2/r)^{-(r+1)/2}$$

where $t = \frac{x - \bar{x}}{s}$ and $\Gamma$ is the gamma function

*Mean* $= 0$ (this applies to all degrees of freedom $r$ except if the distribution is shifted to another nonzero central location)

*Standard Deviation* $= \sqrt{\frac{r}{r-2}}$



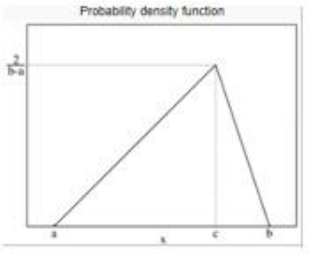
**Distribuição triangular**

- descreve a situação em que os valores mínimo, máximo e mais provável de um evento são conhecidos. Tanto o valor mínimo quanto o máximo são fixos, e o valor mais provável fica entre eles, formando uma distribuição em formato triangular. Por exemplo, essa distribuição pode descrever as vendas de um produto quando conhecemos as estimativas mínimas, máximas e mais prováveis.

$$f(x) = \begin{cases} \dfrac{2(x - Min)}{(Max - Min)(Likely - Min)} & \text{for } Min < x < Likely \\ \dfrac{2(Max - x)}{(Max - Min)(Max - Likely)} & \text{for } Likely < x < Max \end{cases}$$

$$Mean = \frac{1}{3}(Min + Likely + Max)$$

$$Standard\ Deviation = \sqrt{\frac{1}{18}(Min^2 + Likely^2 + Max^2 - MinMax - MinLikely - MaxLikely)}$$



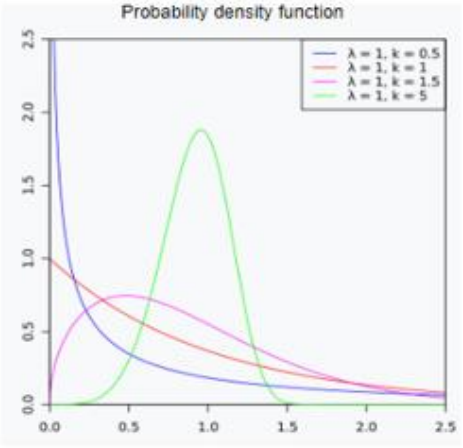
**Distribuição Weibull**

- é muito empregada em testes de fadiga, por exemplo, para descrever o tempo de falha em estudos de confiabilidade ou resistência à ruptura de materiais em testes de controle de qualidade. Também pode ser usado para modelar grandezas físicas, como a velocidade do vento. Depende de 3 parâmetros de entrada: a localização L, a forma alfa e a escala beta. Quando o parâmetro de forma = 1, torna-se a distribuição exponencial.

$$f(x) = \frac{\alpha}{\beta}\left[\frac{x}{\beta}\right]^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$$

$$Mean = \beta\Gamma(1+\alpha^{-1})$$

$$Standard\ Deviation = \beta^2\left[\Gamma(1+2\alpha^{-1}) - \Gamma^2(1+\alpha^{-1})\right]$$



**Distribuição exponencial**

Também é conhecido como uma distribuição exponencial negativa que representa o tempo entre as trilhas em um processo de Poisson. A relação entre a distribuição exponencial e a distribuição de Poisson.

A distribuição exponencial é usada para análise de sobrevivência, por exemplo, vida útil de um ar condicionado, vida útil esperada de uma máquina e período de tempo entre as chegadas ao metrô.

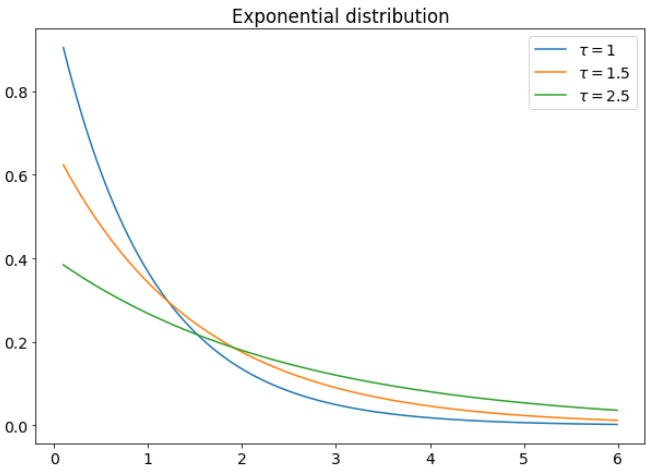
Diz-se que uma variável X possui uma distribuição exponencial quando

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0, \\ 0 & x < 0. \end{cases}$$

Onde λ representa taxa e sempre tem valor maior que zero.

Ou $f(t,\tau) = \frac{1}{\tau} e^{-t/\tau}$ onde τ é o intervalo de tempo médio entre dois eventos consecutivos.

O gráfico de distribuição exponencial é mostrado abaixo;




Algumas das fórmulas disso

• O valor esperado de uma variável aleatória exponencial é dado por:

$$\mathrm{E}[X] = \frac{1}{\lambda}$$

• O valor de variância de uma variável aleatória exponencial é dado por:

$$\mathrm{Var}[X] = \frac{1}{\lambda^2}$$

• A função geradora de momento da variável aleatória exponencial é dada por:

$$M_X(t) = \frac{\lambda}{\lambda - t}$$

• A função característica de uma variável aleatória exponencial é dada por:

$$\varphi_X(t) = \frac{\lambda}{\lambda - it}$$

A distribuição exponencial tem as seguintes características;

- Conforme mostrado no gráfico, quanto maior a taxa, mais rápido a curva cai e, quanto menor a taxa, mais plana a curva.
- Na análise de sobrevivência, λ é denominado como uma taxa de falha de uma máquina em qualquer momento t com a suposição de que a máquina sobreviverá até o momento.

A distribuição exponencial descreve a quantidade de tempo entre eventos que ocorrem em momentos aleatórios. Considera-se que o tempo não tem efeito nos resultados futuros (o tempo de vida futuro de um objeto tem a mesma distribuição, independentemente do tempo em que existiu) o que torna o exponencial "sem memória". Ele pode ser usado para modelar situações como: quanto tempo temos que esperar em uma encruzilhada até ver um carro passando no sinal vermelho ou

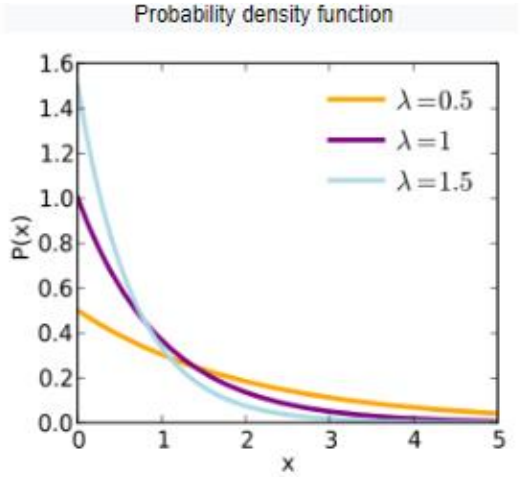
quanto tempo levará até que alguém receba o próximo telefonema? Quanto tempo um produto funcionará antes de quebrar?

A distribuição exponencial está relacionada a Poisson, que não descreve o tempo decorrido, mas o número de ocorrências de um evento em um determinado intervalo de tempo. A distribuição exponencial é parametrizada apenas por lambda, a taxa de sucesso.

$$f(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0; \lambda > 0$$

$$Mean = \frac{1}{\lambda}$$

$$Standard\ Deviation = \frac{1}{\lambda}$$



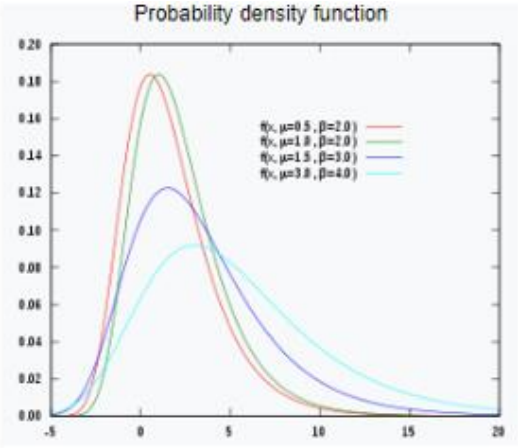
**Distribuição de valores extremos, ou distribuição de Gumbel,**

- modela a distribuição do máximo (ou mínimo) de um número de amostras de várias distribuições. Exemplos dessa distribuição são as resistências à ruptura dos materiais, a carga máxima de uma aeronave, estudos de tolerância, o nível máximo de um rio ou de um terremoto em um determinado ano. Esta distribuição tem 2 parâmetros, o modo m correspondendo ao ponto mais provável (ou o pico mais alto da PDF) e um parâmetro de escala, beta, que é > o e governa a variância.

$$f(x) = \frac{1}{\beta} z e^{-z} \text{ where } z = e^{\frac{x-m}{\beta}} \text{ for } \beta > 0;$$

$$Mean = m + 0.577215\beta$$

$$Standard\ Deviation = \sqrt{\frac{1}{6}\pi^2\beta^2}$$



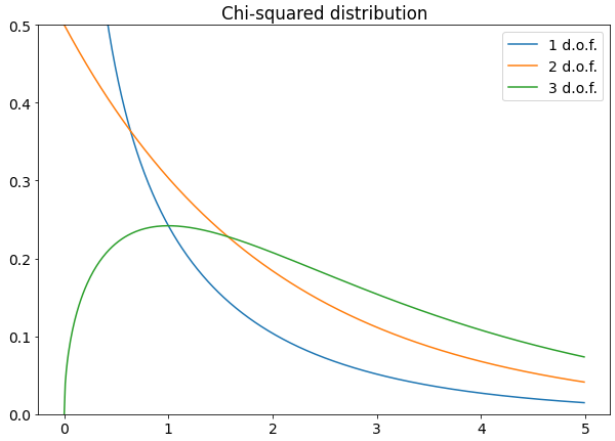
**Distribuição qui-quadrado**

Se você pegar algumas variáveis aleatórias independentes e normalmente distribuídas, pegar seu valor quadrado e somá-los, você obterá uma variável qui-quadrada, cuja função de densidade de probabilidade é:

$$f(x,k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

k é o número de graus de liberdade. Muitas vezes é igual ao número de variáveis normais quadradas que você soma, mas às vezes é reduzido se houver alguma relação entre essas variáveis (por exemplo, o número de parâmetros do modelo em um problema de ajuste de curva).

Aqui estão alguns exemplos dessa distribuição.




Você usará a distribuição qui-quadrado no teste qui-quadrado de Pearson, que compara um histograma experimental com uma distribuição teórica sob certas suposições. A distribuição qui-quadrado está relacionada ao teste F para a igualdade de variâncias porque a distribuição F é calculada como a distribuição de probabilidade da razão entre duas variáveis qui-quadrado.

A distribuição Qui-Quadrado é predominantemente usada em testes de hipóteses, na construção de intervalos de confiança, na avaliação da qualidade de ajuste de uma distribuição observada a uma teórica. A variável qui-quadrado (com um grau de

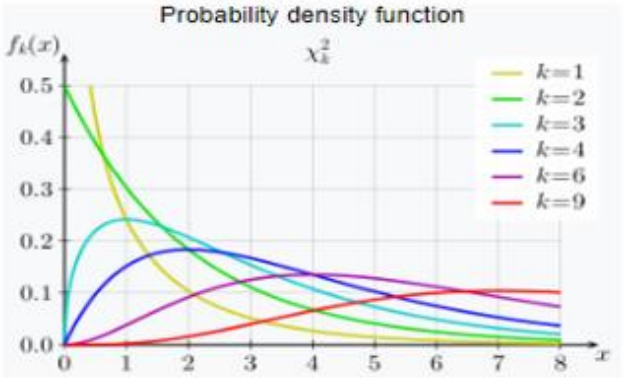
liberdade) é o quadrado de uma variável normal padrão, e a distribuição de qui-quadrado tem propriedade aditiva (a soma de duas distribuições de qui-quadrado independentes também é uma variável de qui-quadrado). A soma de k distribuições normais independentes é distribuída como um qui-quadrado com k graus de liberdade. A distribuição qui-quadrado também pode ser modelada usando uma distribuição gama com o parâmetro de forma como k/2 e escala como $2S^2$.

A distribuição qui-quadrado tem um parâmetro: k, o número de graus de liberdade.

$$f(x) = \frac{2^{(-k/2)}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2} \text{ for all } x > 0$$

$$Mean = k$$

$$Standard\ Deviation = \sqrt{2k}$$



**Distribuição beta**

É a família de distribuições de probabilidade contínua definida no intervalo [0,1], que é expressa por alfa e beta. Além disso, este modelo é usado para o modelo que tem uma incerteza da probabilidade de sucesso de um experimento aleatório. Ele também oferece uma ferramenta poderosa com as estatísticas básicas que podem calcular o nível de confiança do tempo de conclusão.

A distribuição beta vem em distribuições de probabilidade contínuas tendo o intervalo [0,1] com dois parâmetros de forma que podem ser expressos por alfa (α) e beta (β). Esses dois parâmetros são o expoente de uma variável aleatória e controlam a forma da distribuição.

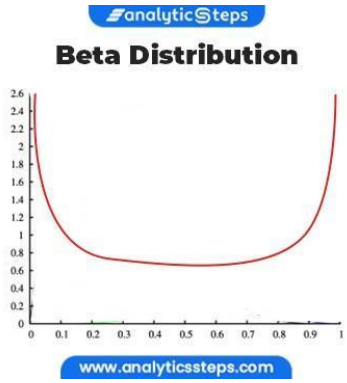
A distribuição mostra a família de probabilidades e é um modelo adequado para representar o comportamento aleatório de porcentagens ou proporções. É usado para os modelos de dados que contêm incertezas das probabilidades de sucesso em um experimento aleatório.

A função de densidade de probabilidade para a distribuição beta é

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

Onde **β** é o segundo parâmetro de forma e B (α, β) é a constante de normalização que garante que a área sob a curva seja um.

O gráfico da distribuição beta é mostrado abaixo;




A formulação geral da distribuição beta também é conhecida como distribuição beta de primeiro tipo e distribuição beta de segundo tipo é outro nome de distribuição beta principal.

A distribuição beta tem muitas aplicações na descrição estatística de frequências de alelos em populações genéticas, alocação de tempo em gerenciamento de projetos, dados de luz solar, proporções de minerais em rochas, etc.

*Propriedades da distribuição beta*

Existem algumas propriedades que podem satisfazer essas distribuições:

Os termos para medir a tendência central são:

- Quer dizer
- Média Harmônica
- Modo
- Mediana
- Média geométrica

Os termos para medir a dispersão estatística são:

- Variância geométrica e covariância
- Variância
- Desvio médio absoluto em torno da média
- Diferença absoluta média

**Distribuição beta-binomial**

É o modelo bayesiano mais simples, amplamente utilizado em testes de inteligência, epidemiologia e marketing. Uma distribuição é considerada beta-binomial se a probabilidade de sucesso for p, e a forma do parâmetro binomial do batimento for α> o e β> o.

A forma paramétrica pode ser definida como a probabilidade de sucesso:

• Uma distribuição pode se aproximar de uma distribuição binomial para o maior valor de α e β.

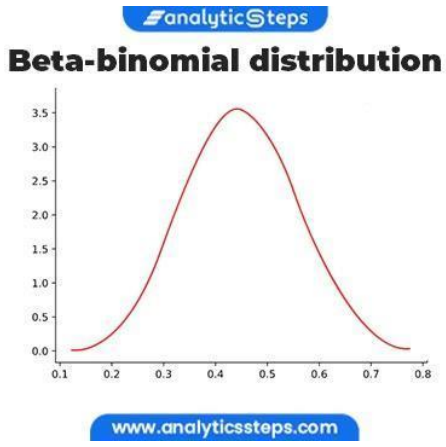
• O valor da distribuição uniforme discreta é igual à distribuição de o a n, se o valor de α e β for igual a 1.

• Para o valor de n = 1, a distribuição beta-binomial é o mesmo valor da distribuição de Bernoulli.

A principal diferença entre uma distribuição beta e uma distribuição binomial é que p é sempre fixo para um conjunto de tentativas em uma distribuição binomial, enquanto o p para beta-binomial não é fixo e muda de trilha a trilha.

Sendo a forma mais simples do modo bayesiano, a distribuição beta-binomial tem amplas aplicações em testes de inteligência, epidemiologia e marketing.

O gráfico da distribuição beta-binomial se parece com o abaixo;



A forma paramétrica pode ser definida na forma de probabilidade de sucesso de tal forma que

- Uma distribuição tende a uma distribuição binomial para o maior valor de α e β.
- O valor da distribuição uniforme discreta é equivalente à distribuição entre o an, se ambos os valores α = β = 1.
- Para n = 1, a distribuição beta-binomial é aproximadamente igual à distribuição de Bernoulli.

Falando sobre a principal diferença entre uma distribuição beta e uma distribuição binomial, a probabilidade de sucesso, p, é sempre fixada para um conjunto de tentativas, ao passo que não é fixada para a distribuição beta-binomial e as mudanças de trilha a trilha.

A distribuição beta é comumente usada para representar a variabilidade em um intervalo fixo. Por exemplo, para modelar o comportamento de variáveis aleatórias limitadas a intervalos de comprimento finito. Também é uma escolha adequada para modelar porcentagens ou proporções. Na inferência Bayesiana, ela é usada para modelar a priori conjugada para distribuições de Bernoulli, Binomial, Binomial Negativa e geométrica. Simplificando, a distribuição beta é uma boa proposta para os priors (o conhecimento inicial do sucesso) para diferentes aplicações da família Bernoulli, como o número de caras em tentativas de lançamento de moedas ou qualquer outro evento de resultado duplo. Leva 2 parâmetros, alfa e beta, e a variável incerta é um valor aleatório entre o e um valor positivo. Diferentes combinações de alfa e beta levam às seguintes formas de distribuição:

alfa == beta => distribuição simétrica

if (alfa == 1 e beta > 1) ou (beta== 1 e alfa> 1) => distribuição em forma de J

alfa < beta => desvio positivo

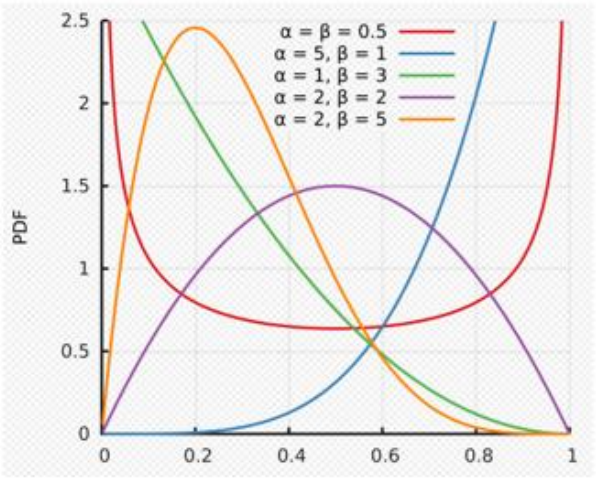
alfa >beta => desvio negativo

$$f(x) = \frac{(x)^{(\alpha-1)}(1-x)^{(\beta-1)}}{\left[\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right]} \text{ for } \alpha > 0;\ \beta > 0;\ x > 0$$

$$Mean = \frac{\alpha}{\alpha+\beta}$$

$$Standard\ Deviation = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}}$$

$$\Gamma(n) = (n-1)!$$



**Distribuição gama**

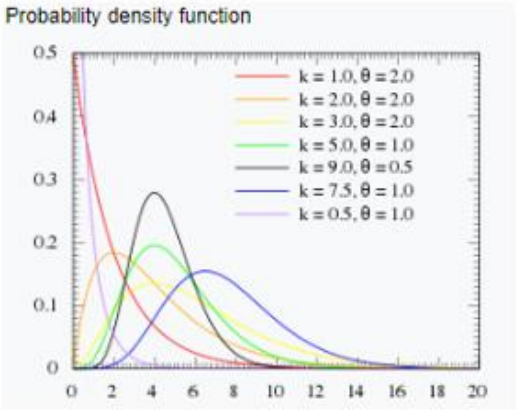
- usada para medir o tempo entre a ocorrência de eventos quando o processo do evento não é completamente aleatório. O número de eventos no período estudado não se limita a um número fixo. Os eventos são independentes. A distribuição gama está relacionada às distribuições lognormal, exponencial Pascal, Poisson e qui-quadrado. Ele pode ser usado para modelar concentrações de poluentes e quantidades de precipitação em processos meteorológicos. Depende de 2 parâmetros, alfa ou parâmetro de forma e beta, parâmetro de escala.

  Um caso especial surge quando alfa é um número inteiro positivo, caso em que a distribuição (também conhecida como distribuição Erlang) pode ser usada para prever tempos de espera em sistemas de filas.

$$f(x) = \frac{\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta} \text{ with any value of } \alpha > 0 \text{ and } \beta > 0$$

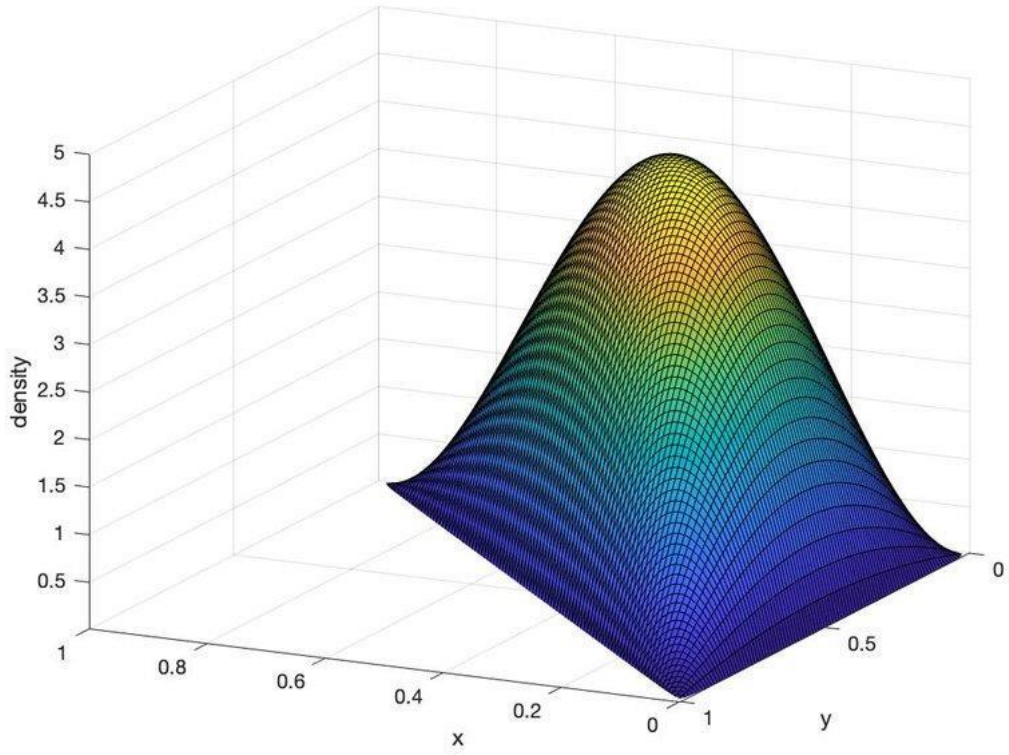
$$Mean = \alpha\beta$$

$$Standard\ Deviation = \sqrt{\alpha\beta^2}$$



**Distribuição de Dirichlet**

- uma generalização multivariada das distribuições beta, razão pela qual também é conhecida como distribuições beta multivariadas. É usado como distribuição a priori na estatística Bayesiana, onde é a priori conjugada da distribuição categórica e multinomial. É parametrizado por um vetor alfa de reais positivos, e amostra sobre um simplex de probabilidade. Um simplex de probabilidade é um conjunto de k números somados a 1 e que correspondem às probabilidades de k classes. Uma distribuição de Dirichlet k-dimensional tem k parâmetros.

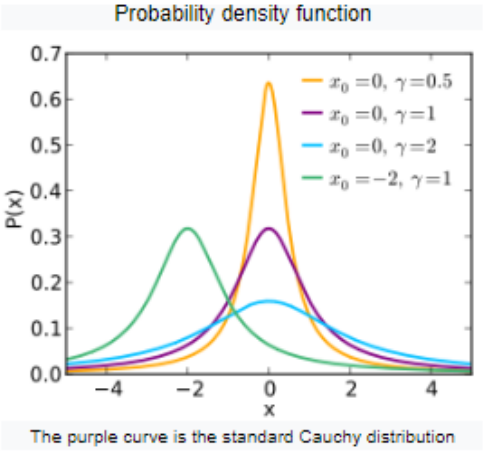


**Distribuição de Cauchy**

- empregada em teoria mecânica e elétrica, antropologia física e medição e problemas de calibração. Na física, descreve a distribuição da energia de um estado instável na mecânica quântica sob o nome de distribuição Lorentziana. Outra aplicação é modelar os pontos de impacto de uma linha reta fixa de partículas emitidas de uma fonte pontual ou em estudos de robustez. A distribuição de Cauchy é conhecida por ser uma distribuição patológica, pois sua média e variância são indefinidas. Leva dois parâmetros. Na estatística Bayesiana, a distribuição de Cauchy pode ser usada para modelar as prioris para os coeficientes de regressão na regressão logística.

  A distribuição tem dois parâmetros, o modo m (correspondente ao pico) e a escala gama (meia largura na metade do máximo da distribuição). A distribuição de Cauchy é a distribuição T de Student com 1 grau de liberdade.

$$f(x) = \frac{1}{\pi} \frac{\gamma/2}{(x-m)^2 + \gamma^2/4}$$



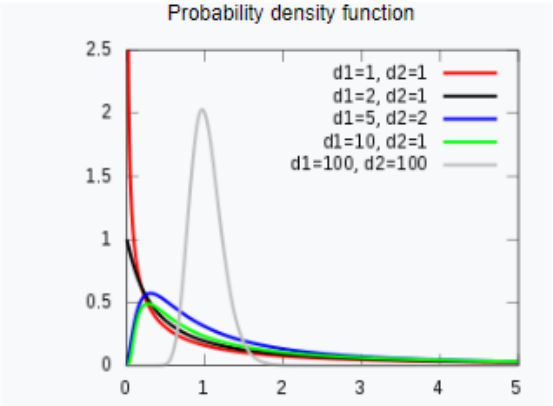
**Distribuição F**

- usada para testar a diferença estatística entre duas variâncias como parte de uma análise ANOVA unidirecional ou a significância geral de um modelo de regressão com testes f. Freqüentemente é a distribuição nula (a distribuição de probabilidade quando a hipótese nula é verdadeira) das estatísticas de teste. Toma como parâmetros n graus de liberdade para o numerador e m graus de liberdade para o denominador.

$$\frac{\chi_n^2/n}{\chi_m^2/m} \overset{d}{\sim} F_{n,m} \text{ or } f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)\left(\frac{n}{m}\right)^{n/2} x^{n/2-1}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)\left[x\left(\frac{n}{m}\right)+1\right]^{(n+m)/2}}$$

$$Mean = \frac{m}{m-2}$$

$$Standard\ Deviation = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)} \text{ for all } m > 4$$

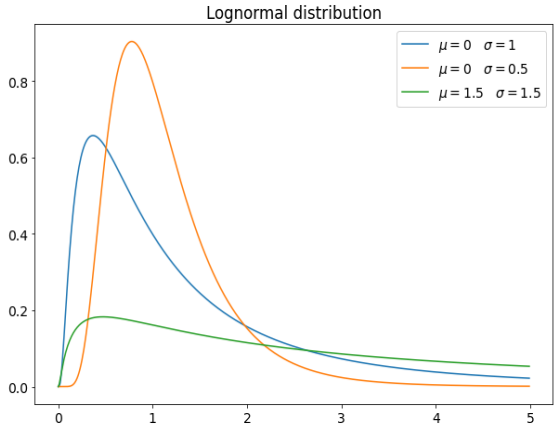


**Distribuição logarítmica**

Se o log da potência for distribuído normalmente, a variável será considerada como log-normalmente distribuída. Ou podemos dizer que ln (x) é normalmente distribuído e que a variável x é assumida como tendo uma distribuição log-normal.

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

μ e σ são os mesmos da distribuição de Gauss.




*Propriedades da distribuição log-normal*

• O valor esperado ou a média de distribuição oferece dados úteis sobre o que uma média esperaria de um número de trilha repetido.

• A mediana de uma distribuição log-normal é outra consideração da tendência central e é útil para outliers que ajudam os meios a liderar.

• O modo de distribuição é um valor com a maior probabilidade de ocorrer.

• Como espalhar a informação pode ser medido pela variância. A raiz quadrada da variância e o desvio padrão são úteis porque têm a mesma unidade dos dados.

Esses valores são muito mais fáceis de medir para uma distribuição de probabilidade contínua. Mas como sua medida inclui uma boa quantidade de cálculo, a descrição pode ser breve.

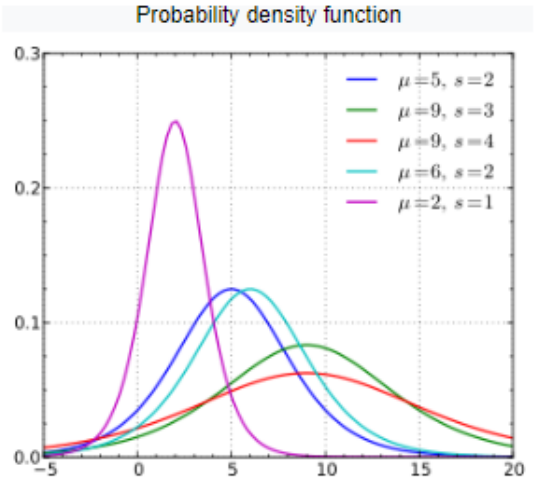
**Distribuição logística**

- usada para descrever o crescimento populacional ao longo do tempo ou reações químicas. Esta distribuição é simétrica e toma 2 parâmetros: a média ou o valor médio e a escala, controlando a variância.

$$f(x) = \frac{e^{\frac{\mu - x}{\alpha}}}{\alpha\left[1 + e^{\frac{\mu - x}{\alpha}}\right]^2} \text{ for any value of } \alpha \text{ and } \mu$$

$$Mean = \mu$$

$$Standard\ Deviation = \sqrt{\frac{1}{3}\pi^2\alpha^2}$$



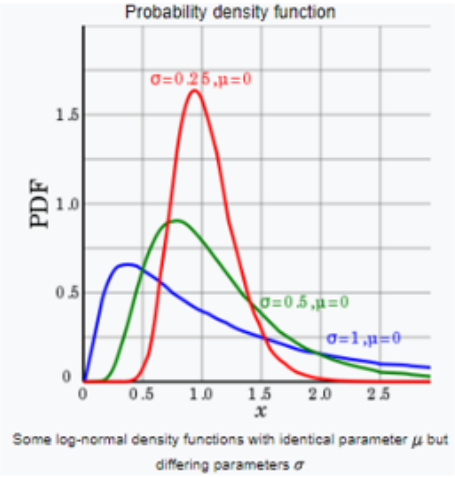
Distribuição lognormal

- uma boa candidata para modelar valores positivamente assimétricos, que são ≥ 0. Por exemplo, a distribuição normal não pode ser usada para modelar os preços das ações porque tem um lado negativo e os preços das ações não podem cair abaixo de zero, então a distribuição lognormal é um bom candidato. Assim, se uma variável aleatória X é log-normalmente distribuída, então Y = ln(X) tem uma distribuição normal. Da mesma forma, se Y tem uma distribuição normal, então a função exponencial de Y, X = exp(Y), tem uma distribuição lognormal. A distribuição lognormal é descrita por 2 parâmetros, a média e o desvio padrão.

$$f(x) = \frac{1}{x\sqrt{2\pi\ln(\sigma)}} e^{-\frac{[\ln(x)-\ln(\mu)]^2}{2[\ln(\sigma)]^2}} \text{ for } x > 0; \mu > 0 \text{ and } \sigma > 0$$

$$Mean = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

$$Standard\ Deviation = \sqrt{\exp(\sigma^2 + 2\mu)[\exp(\sigma^2) - 1]}$$



Some log-normal density functions with identical parameter $\mu$ but differing parameters $\sigma$

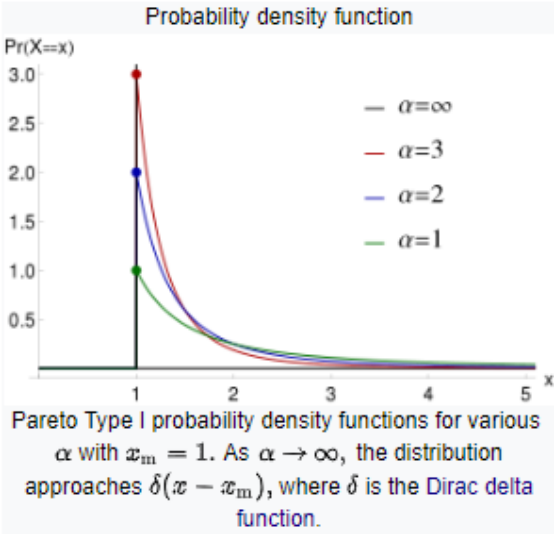
**Distribuição de Pareto**

- uma distribuição de probabilidade de lei de potência usada para modelar fenômenos empíricos como a distribuição de riqueza, as flutuações dos preços das ações, a ocorrência de recursos naturais. A distribuição de Pareto foi vulgarizada sob o nome de princípio de Pareto (ou a "regra 80-20", o princípio de Mateus) afirmando que, por exemplo, 80% da riqueza de uma sociedade é detida por 20% de sua população. No entanto, a distribuição de Pareto só produz este resultado para um determinado valor de potência do parâmetro de entrada alfa (α = log45 ≈ 1,16). Em termos de parâmetros, essa distribuição depende da localização (o limite inferior da variável) e da forma que controla a variância.

$$f(x) = \frac{\beta L^{\beta}}{x^{(1+\beta)}} \text{ for } x > L$$

$$Mean = \frac{\beta L}{\beta - 1}$$

$$Standard\ Deviation = \sqrt{\frac{\beta L^2}{(\beta-1)^2(\beta-2)}}$$




Pareto Type I probability density functions for various $\alpha$ with $x_m = 1$. As $\alpha \to \infty$, the distribution approaches $\delta(x - x_m)$, where $\delta$ is the Dirac delta function.

**Distribuições discretas**

Bernoulli

Binomial

Geometric

Hypergeometric

Negative binomial

Discrete uniform

Poisson

**Distribuição Bernoulli**

Uma distribuição de Bernoulli é um tipo de distribuição de probabilidade discreta - uma tentativa aleatória que tem dois resultados. Existe um caso especial que possui o valor n = 1, por exemplo, um único lançamento de moeda.

Esta é uma das distribuições mais simples que pode ser usada como um ponto inicial para derivar distribuições mais complexas. A distribuição de Bernoulli tem possivelmente dois resultados (sucesso ou fracasso) e uma única tentativa.

Por exemplo, jogando uma moeda, a probabilidade de sucesso de um resultado ser cara é p, então a probabilidade de ter uma cauda como resultado é (1-p). A distribuição de Bernoulli é o caso especial da distribuição binomial com um único ensaio.

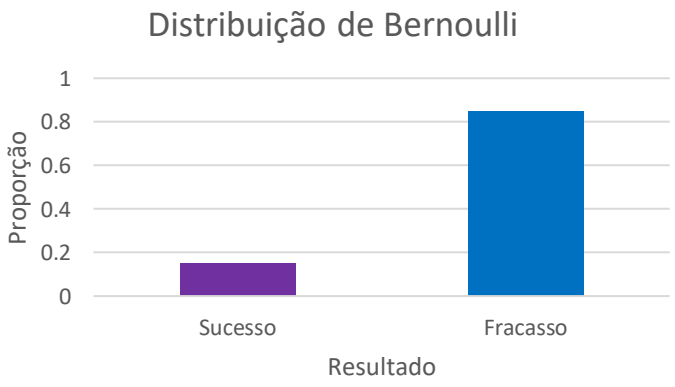
A função de densidade pode ser dada como

f(x) = px (1-p)(1-x) where x ∈ (0,1)

Também pode ser escrito como;

$$P(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

O gráfico da distribuição de Bernoulli é mostrado abaixo, onde a probabilidade de sucesso é menor do que a probabilidade de fracasso.



*Características da distribuição Bernoulli*

• O número de tentativas que devem ser realizadas em um único experimento deve ser predefinido.

• Cada trilha deve ter dois resultados que sejam sucesso ou fracasso.

• A probabilidade de sucesso em cada experimento deve ser a mesma.

• O experimento deve ser independente um do outro, o que significa que o resultado de um ensaio não é afetado pelo resultado do outro.

*Propriedades*

O valor esperado da variável selecionada aleatoriamente é dado por E (x) = p e pode ser derivado:

E(x) = 0*(1-p) + 1 * p = p

A variância da variável Bernoulli é dada por p * (1-p) e é dada como:

Var(X) = p – p2 = p*(1-p)

**Distribuição binomial**

É uma distribuição de probabilidade que conclui o valor que assume um de dois valores independentes sob um conjunto de suposições ou parâmetros. Além disso, as suposições da distribuição binomial devem ter um único resultado com a mesma probabilidade de sucesso. E essa trilha deve ser independente uma da outra.

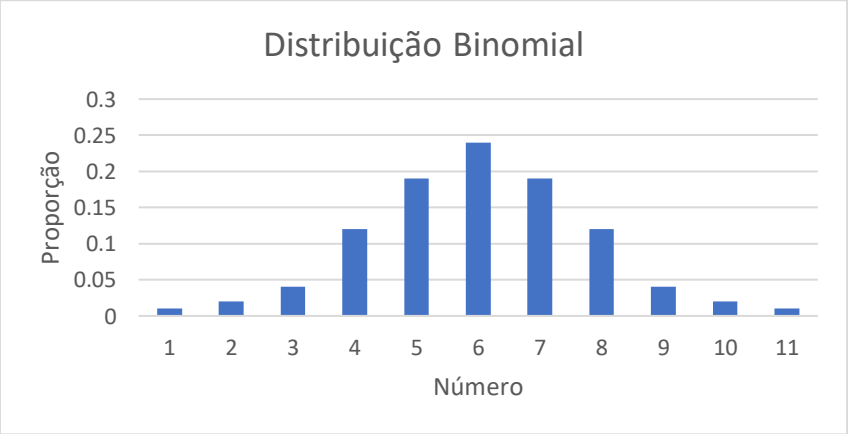
A distribuição binomial é aplicada em eventos de resultados binários onde a probabilidade de sucesso é igual à probabilidade de falha em todas as tentativas sucessivas. Seu exemplo inclui o lançamento de uma moeda tendenciosa / imparcial várias vezes.

Como entrada, a distribuição considera dois parâmetros e, portanto, é chamada de distribuição bi-paramétrica. Os dois parâmetros são: o número de vezes que um evento ocorre, n e probabilidade atribuída, p, a uma das duas classes

Para n número de tentativas e probabilidade de sucesso, p, a probabilidade de evento bem-sucedido (x) dentro de n tentativas pode ser determinada pela seguinte fórmula

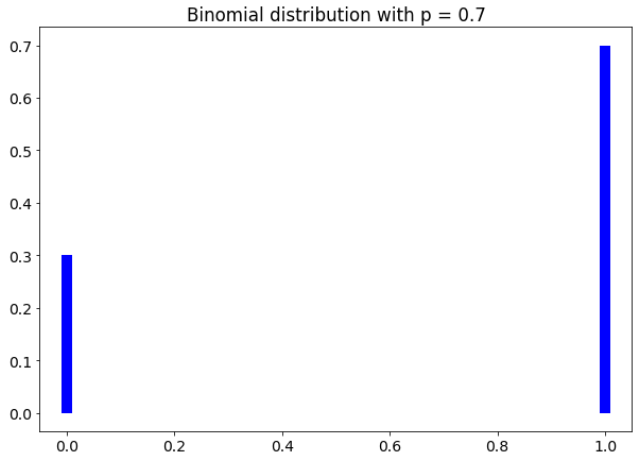
$$P\left(X = x\right) = \frac{n!}{x!(n-x)!}\, p^x \left(1-p\right)^{(n-x)}$$

O gráfico da distribuição binomial é mostrado abaixo quando a probabilidade de sucesso é igual à probabilidade de fracasso.



Se nossos eventos são números 0 e 1 e o evento 1 ocorre com probabilidade p, podemos facilmente escrever a densidade usando a distribuição delta de Dirac:

Uma possível representação é a seguinte:




*Propriedades de uma distribuição binomial*

• Quando um experimento tem trilhas independentes e cada uma delas tem dois resultados que são sucesso e fracasso.

• A distribuição binomial também é chamada de distribuição bi-paramétrica. Como é classificado por dois parâmetros n e p.

Cada uma das tentativas pode ter dois resultados possíveis, sucesso ou fracasso, com probabilidades p e (1-p). Um número total de n tentativas idênticas pode ser conduzido, e a probabilidade de sucesso e falha é a mesma para todas as tentativas.

• O valor médio disso é:

$\mu = np$

• A variância da distribuição binomial é dada por:

$\sigma^2 = npq$

• O valor de p e q é sempre menor ou igual a 1, ou podemos dizer que a variância deve ser menor que seu valor médio.

$npq < np$

**Distribuição Multinomial**

A distribuição multinomial é usada para medir os resultados de experimentos que têm duas ou mais variáveis. É o tipo especial de distribuição binomial quando há dois resultados possíveis, como verdadeiro / falso ou sucesso / falha.  A distribuição é comumente usada em aplicações biológicas, geológicas e financeiras.

Um experimento de Mendel muito popular em que duas linhagens de ervilhas (uma semente verde e enrugada e outra amarela e lisa) são hibridizadas que produziu quatro variedades diferentes de sementes - verde e enrugada, verde e redonda, amarela e redonda, e amarela e enrugada. Isso resultou na distribuição multinomial e levou à descoberta dos princípios básicos da genética.

A função de densidade para distribuição multinomial é

$$P = \frac{n!}{(n_1!)(n_2!)\ldots(n_x!)} P_1{}^{n1} P_2{}^{n2} \ldots P_x{}^{nx}$$

Onde n = número de experimento; Px = probabilidade de ocorrência de um experimento.

O gráfico de distribuição exponencial é mostrado abaixo


A seguir estão as propriedades da distribuição multinomial;

- Um experimento pode ter um número repetido de tentativas, por exemplo, jogar um dado várias vezes.

- Cada tentativa é independente uma da outra.
- A probabilidade de sucesso de cada resultado deve ser a mesma (constante) para todas as tentativas de um experimento.

**Distribuição Poisson**

É uma ferramenta que serve para prever uma determinada probabilidade do evento quando você sabe o valor de acontecer de um determinado evento. A distribuição de Poisson nos fornece a probabilidade de um número aplicado de eventos ocorrer em um período de tempo fixo.

Sendo uma parte da distribuição de probabilidade discreta, a distribuição de poisson descreve a probabilidade de um determinado número de eventos que ocorrem em um período de tempo ou espaço fixo, ou intervalos particularizados, como distância, área, volume.

Por exemplo, a realização de análises de risco por parte do setor segurador / bancário, antecipando o número de acidentes de trânsito em um determinado intervalo de tempo e em uma determinada área.

A distribuição de Poisson considera as seguintes suposições;

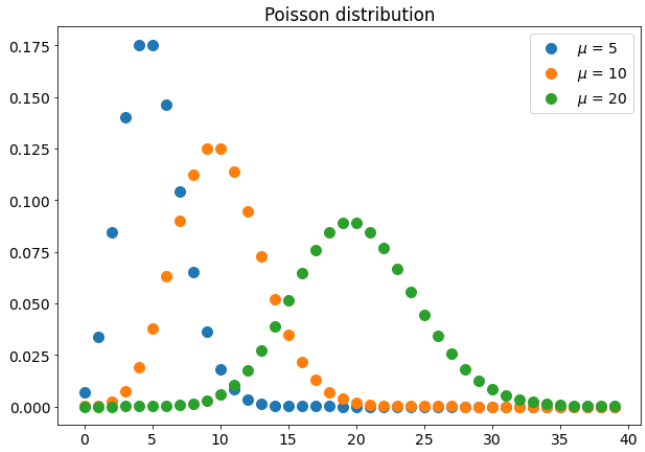
- A probabilidade de sucesso para um curto período é igual à probabilidade de sucesso para um longo período de tempo.
- A probabilidade de sucesso em uma duração é igual a zero conforme a duração se torna menor.
- Um evento de sucesso não pode afetar o resultado de outro evento de sucesso

Uma distribuição de Poisson pode ser modelada usando a fórmula abaixo,

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

Onde **λ** representa o número possível de eventos que ocorrem em um período de tempo fixo e X é o número de eventos naquele período de tempo.

O gráfico da distribuição de Poisson é mostrado abaixo;




*Propriedades da distribuição de Poisson*

• O valor esperado e a variação da variável aleatória são equivalentes a λ.

• O desvio absoluto está associado à média é dado por:

$$E[X-\lambda] = 2\exp(-\lambda)\frac{\lambda^{(\lambda)+1}}{[\lambda]!}$$

- Para a distribuição de Poisson, a média é igual à variância $\sigma^2$ de modo que o CV deve ser $\sigma / \mu = \sigma / \sigma^2 = 1 / \sigma$
- O valor esperado da distribuição de Poisson é decomposto pelo produto subjacente de intensidade e exposição.
- A média da distribuição de Poisson é dada por "m".

- Os eventos são independentes uns dos outros, ou seja, se um evento ocorrer, ele não afeta a probabilidade de outro evento ocorrer.
- Um evento pode ocorrer qualquer número de vezes em um período de tempo definido.
- Dois eventos quaisquer não podem estar ocorrendo ao mesmo tempo.

- A taxa média de eventos que ocorrem é constante.

Distribuição de Poisson

$$P(x) = (e^{-\lambda} * \lambda^{x}) / x!$$

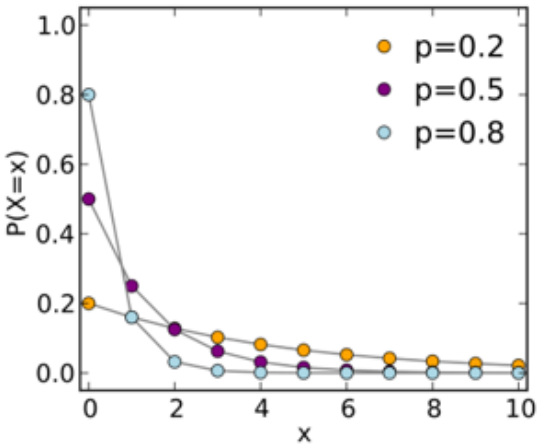
**A distribuição geométrica**

calcula o número de tentativas antes que ocorra o primeiro sucesso. O número de tentativas não é fixo, então a probabilidade de sucesso é a mesma em todas as tentativas independentes, e o experimento continua até o primeiro sucesso. Esta distribuição tem um parâmetro, a probabilidade de sucesso p.

$$P(x) = p(1-p)^{x-1} \text{ for } 0 < p < 1$$

$$Mean = \frac{1}{p} - 1$$

$$Standard\ Deviation = \sqrt{\frac{1-p}{p^2}}$$

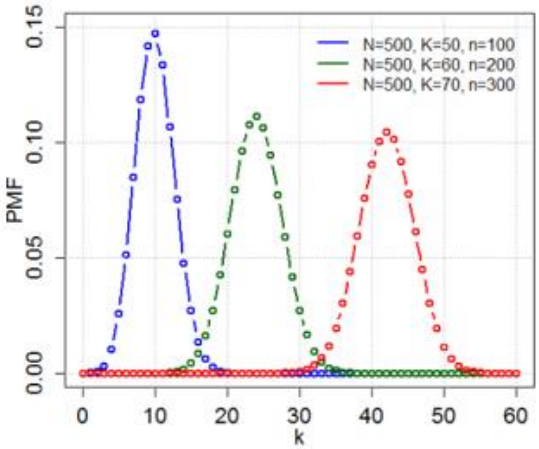



**A distribuição hipergeométrica**

mede o número de sucessos em n tentativas (semelhante ao binomial), mas não se baseia na suposição de independência entre as tentativas. Assim, as tentativas são realizadas sem reposição, e cada tentativa altera a probabilidade de sucesso. Exemplos são a probabilidade de tirar uma certa combinação de cartas de um baralho sem reposição ou selecionar lâmpadas defeituosas de um caixote com lâmpadas defeituosas e funcionando. Essa distribuição depende do número de itens na população (N), do número de tentativas amostradas (n) e do número de itens na

população com a característica de sucesso Nx. x representa o número de tentativas bem-sucedidas.

$$P(x) = \frac{\frac{(N_x)!}{x!(N_x - x)!}\frac{(N - N_x)!}{(n-x)!(N - N_x - n + x)!}}{\frac{N!}{n!(N-n)!}}$$



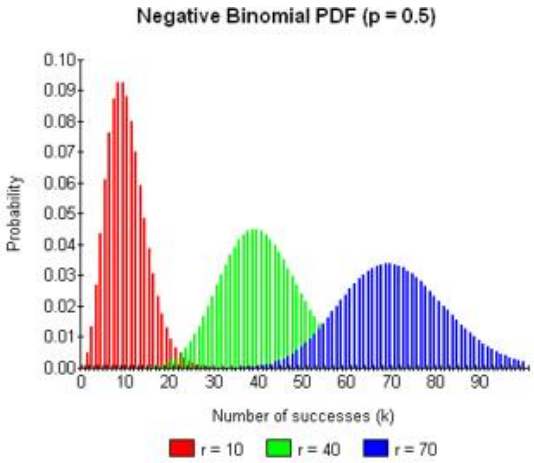
## O binômio negativo

calcula o número de tentativas até atingir r eventos bem-sucedidos. É uma superdistribuição da distribuição geométrica e pode ser usada para modelar situações como o número de visitas de vendas a serem realizadas para fechar r negócios. Os parâmetros usados por esta distribuição são a probabilidade de sucesso p e o número de sucessos requeridos r.

$$P(x) = \frac{(x+r-1)!}{(r-1)!x!}p^r(1-p)^x \text{ for } x = r, r+1, \ldots$$

$$Mean = \frac{r(1-p)}{p}$$

$$Standard\ Deviation = \sqrt{\frac{r(1-p)}{p^2}}$$




## Distribuição uniforme

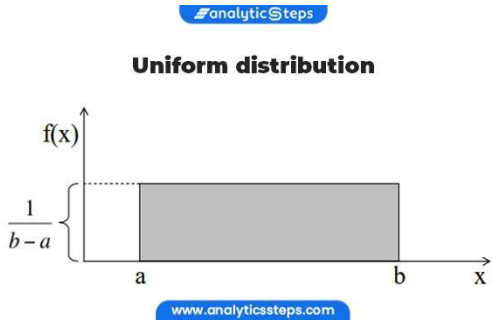
A forma básica de uma distribuição contínua é conhecida como distribuição uniforme. Tem a probabilidade constante de formar uma distribuição retangular. E isso implica que cada valor tem o mesmo comprimento de distribuição. Que têm igual probabilidade de ocorrência. Em contraste, esta função pertence ao tipo de distribuição de probabilidade de entropia máxima.

A distribuição uniforme pode ser discreta ou contínua, onde cada evento é igualmente provável de ocorrer. Tem uma probabilidade constante de construir uma distribuição retangular.

Nesse tipo de distribuição, um número ilimitado de resultados será possível e todos os eventos têm a mesma probabilidade, semelhante à distribuição de Bernoulli. Por exemplo, ao lançar um dado, os resultados são de 1 a 6, que têm probabilidades iguais de ⅙ e representam uma distribuição uniforme.

Diz-se que uma variável X tem distribuição uniforme se a função de densidade de probabilidade é

$$f(x) = \frac{1}{b-a} \qquad for\ -\infty < a \le x \le b < \infty$$




*Características de distribuição uniforme*

• A função de densidade combina para a unidade.

• Cada uma de suas funções de entrada tem peso igual.

• A média da função uniforme é dada por:

$$\mu = \frac{(a+b)}{2}$$

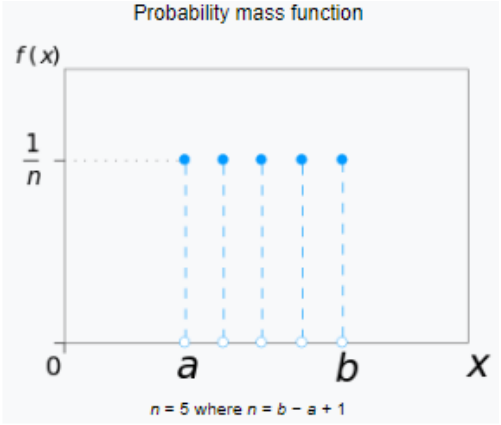
• A variância da distribuição uniforme é dada por:

$$V(x) = \frac{(b-a)^2}{12}$$

$$P(x) = \frac{1}{N} \text{ ranked value}$$

$$Mean = \frac{N+1}{2} \text{ ranked value}$$

$$Standard\ Deviation = \sqrt{\frac{(N-1)(N+1)}{12}}$$

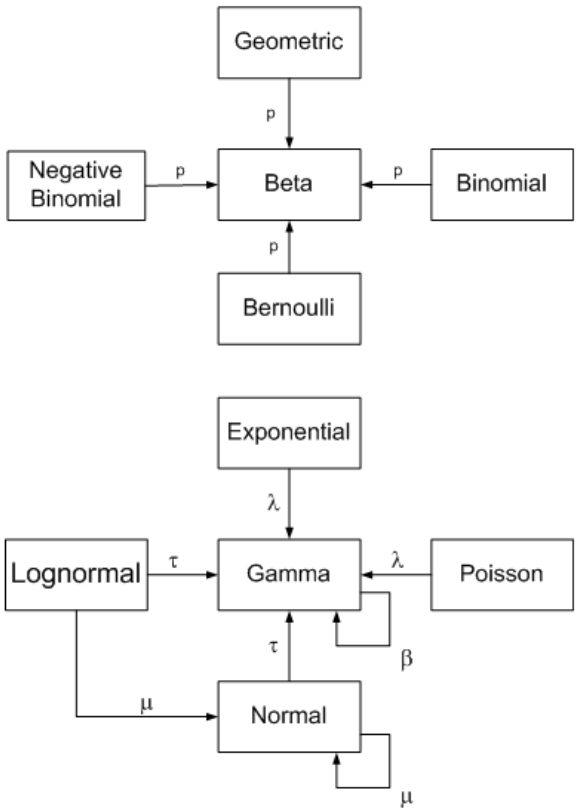


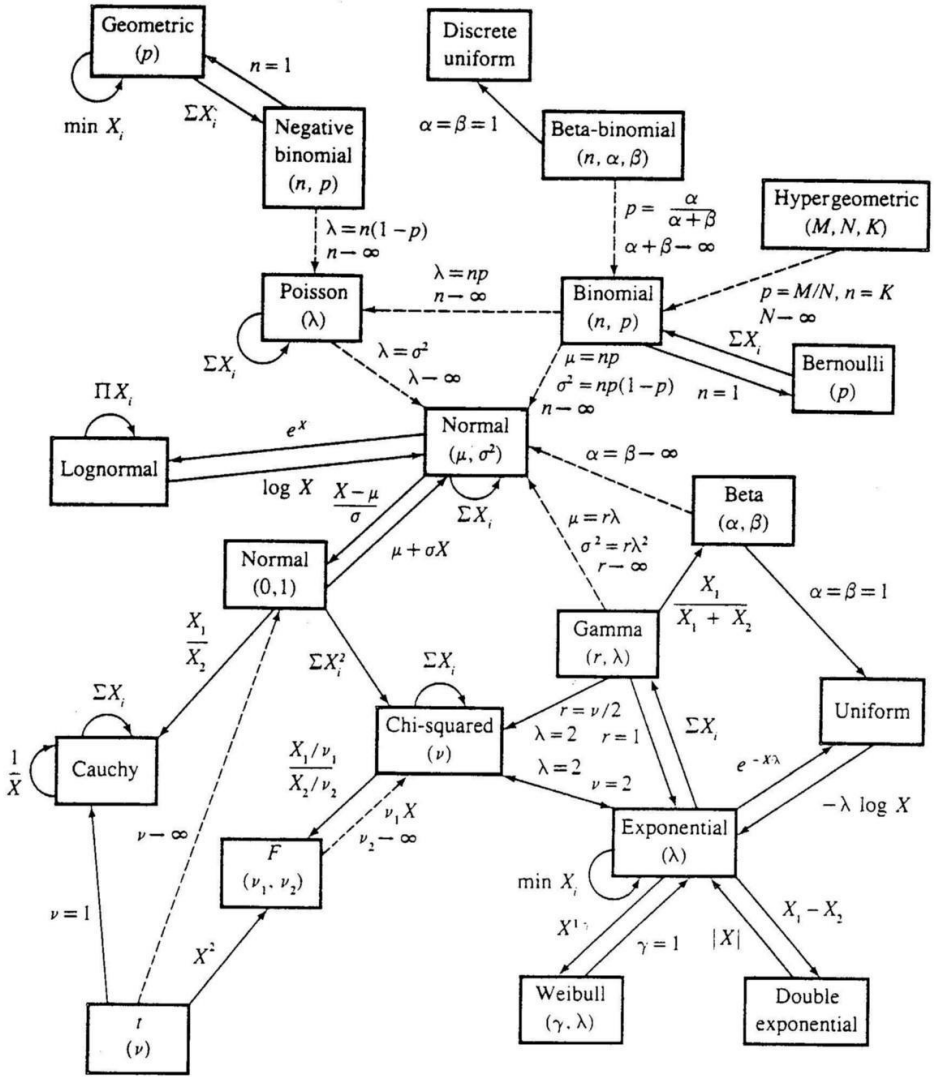
**Distribuições conjugadas**

No contexto da análise Bayesiana, se a distribuição posterior p(teta|x) e a anterior p(teta) fazem parte da mesma família de probabilidades, elas são chamadas de distribuições conjugadas. Além disso, o prior chamou o prior conjugado para a função de verossimilhança.

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')d\boldsymbol{\theta}'}$$

Diferentes escolhas de priori podem tornar a integral mais ou menos difícil de calcular. Se a verossimilhança p(x|teta) tem a mesma forma algébrica que a anterior, podemos obter uma expressão de forma fechada para a posterior.






**Relationships among common distributions**. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

4

[4] https://www.kdnuggets.com/2020/06/overview-data-distributions.html

**Transformação de dados**[5]

Muitas variáveis biológicas não atendem aos pressupostos dos testes estatísticos paramétricos: não são normalmente distribuídas, os desvios padrão não são homogêneos ou ambos. O uso de um teste estatístico paramétrico (como uma anova ou regressão linear) nesses dados pode dar um resultado enganoso. Em alguns casos, transformar os dados fará com que eles se ajustem melhor às suposições.

A transformação de dados é um conceito que se refere à função matemática aplicada a cada valor no conjunto de dados para substituir o valor em um novo valor. Em uma equação matemática, poderíamos expressá-la na imagem abaixo.

$$y_i = f(x_i)$$

Onde $Y_i$ é o dado transformado, f é a função da transformação e $x_i$ é o dado original.

**Por que precisamos fazer Transformação de Dados?**
Existe algum benefício em transformar dados?

Do ponto de vista estatístico, as razões são:

- A transformação de dados permitiu que você cumprisse certas suposições estatísticas, por exemplo, Normalidade, Homogeneidade, Linearidade, etc.
- A transformação de dados dimensiona os valores de diferentes colunas para serem comparáveis, por exemplo, Salário em USD (intervalo de 100 a 10.000) com Peso em quilogramas (intervalo de 20 a 100).
- A transformação de dados é útil para obter novos insights e eliminar ruídos em seus dados. No entanto, utilizar o método de transformação de dados exigia que você entendesse o efeito, a implicação e a conclusão da transformação com base nos dados transformados. Na minha opinião, você só faz transformação de dados se for necessário e entender seu objetivo de transformação.

[5]https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Biological_Statistics_(McDonald)/04%3A_Tests_for_One_Measurement_Variable/4.06%3A_Data_Transformations

- Melhorar a interpretabilidade. Algumas variáveis não estão no formato que precisamos para uma determinada pergunta, por exemplo, os fabricantes de automóveis fornecem valores de milhas/galões para consumo de combustível, no entanto, para comparar modelos de carros, estamos mais interessados nos galões/milhas recíprocos.
- Gráficos de desorganização. Se você visualizar duas ou mais variáveis que não estão distribuídas uniformemente pelos parâmetros, você acabará com pontos de dados próximos. Para uma melhor visualização, pode ser uma boa ideia transformar os dados para que sejam distribuídos de maneira mais uniforme no gráfico. Outra abordagem pode ser usar uma escala diferente no eixo do gráfico.
- Para obter informações sobre a relação entre as variáveis. A relação entre variáveis muitas vezes não é linear, mas de um tipo diferente. Um exemplo comum é usar o logaritmo da renda para compará-lo com outra variável, pois a utilidade de mais renda diminui com a renda mais alta. (Veja esta excelente discussão sobre a transformação logarítmica altamente utilizada na validação cruzada.) Outro exemplo é o crescimento polinomial de dinheiro em uma conta bancária com taxa de juros comparada ao tempo. Para calcular um coeficiente de correlação simples entre variáveis, as variáveis precisam mostrar uma relação linear. Para atender a esses critérios, você pode transformar uma ou ambas as variáveis.
- Atender aos pressupostos para inferência estatística. Ao construir intervalos de confiança simples, a suposição é que os dados sejam normalmente distribuídos e não enviesados para a esquerda ou para a direita. Para a análise de regressão linear, uma suposição importante é a homocedasticidade, o que significa que a variância do erro de sua variável de resultado dependente é independente de suas variáveis de previsão. Uma suposição para muitos testes estatísticos como o teste T é que os erros de um modelo (os valores de uma medida amostrada de uma população) são normalmente distribuídos.
-

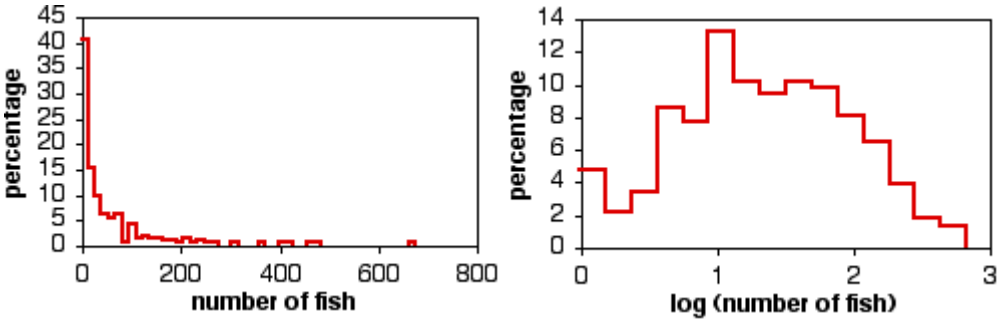



Fig.1 Histogramas do número de mudminnows orientais por seção de 75 m de fluxo (amostras com o mudminnows excluídos). Dados não transformados à esquerda, dados transformados em log à direita.

Para transformar dados, você executa uma operação matemática em cada observação e, em seguida, usa esses números transformados em seu teste estatístico. Por exemplo, como mostrado no primeiro gráfico acima, a abundância da espécie de peixe *Umbra pygmaea* (Eastern mudminnow) nos riachos de Maryland não tem distribuição normal; há muitos córregos com uma pequena densidade de mudminnows, e alguns córregos com muitos deles. A aplicação da transformação de log torna os dados mais normais, conforme mostrado no segundo gráfico.

Fig. 2 Eastern mudminnow

Aqui estão 12 números do conjunto de dados do mudminnow; a primeira coluna são os dados não transformados, a segunda coluna é a raiz quadrada do número na primeira coluna e a terceira coluna é o logaritmo de base 10 do número na primeira coluna.

| **Untransformed** | **Square-root** | **Log** |
|---|---|---|
| | transformed | transformed |
| **38** | 6.164 | 1.58 |

| | | |
|---|---|---|
| **1** | 1 | 0 |
| **13** | 3.606 | 1.114 |
| **2** | 1.414 | 0.301 |
| **13** | 3.606 | 1.114 |
| **20** | 4.472 | 1.301 |
| **50** | 7.071 | 1.699 |
| **9** | 3 | 0.954 |
| **28** | 5.292 | 1.447 |
| **6** | 2.449 | 0.778 |
| **4** | 2 | 0.602 |
| **43** | 6.557 | 1.633 |

Você faz as estatísticas sobre os números transformados. Por exemplo, a média dos dados não transformados é 18,9; a média dos dados transformados em raiz quadrada é 3,89; a média dos dados log transformados é 1,044. Se você estivesse comparando a abundância de peixes em diferentes bacias hidrográficas e decidisse que a transformação logarítmica era a melhor, faria uma anova unidirecional nos logaritmos da abundância de peixes e testaria a hipótese nula de que as médias do logaritmo. as abundâncias transformadas eram iguais.

**Quais são os métodos para transformação de dados?**

De acordo com McCune e Grace (2002) em seu livro Análise de Comunidades Ecológicas, os métodos são:

Transfromação Monotônica

Relativizações (Padronização)

Transformação probabilística (suavização)

**Transformação de volta**

Mesmo que você tenha feito um teste estatístico em uma variável transformada, como o log da abundância de peixes, não é uma boa ideia relatar suas médias, erros padrão, etc. em unidades transformadas. Um gráfico que mostrasse que a média do logaritmo de peixes por 75m de riacho era de 1,044 não seria muito informativo para quem não sabe fazer expoentes fracionários de cabeça. Em vez disso, você deve transformar seus resultados de volta. Isso envolve fazer o oposto da função

matemática que você usou na transformação de dados. Para a transformação de log, você transformaria de volta elevando 10 à potência do seu número. Por exemplo, os dados log transformados acima têm uma média de 1,044 e um intervalo de confiança de 95% de ±0,344 peixes transformados em log. A média retransformada seria 101,044=11,1 peixes. O limite de confiança superior seria 10(1,044+0,344)=24,4 peixes, e o limite de confiança inferior seria 10(1,044-0,344)=5,0 peixes. Observe que o intervalo de confiança não é simétrico; o limite superior é de 13,3 peixes acima da média, enquanto o limite inferior é de 6,1 peixes abaixo da média. Observe também que você não pode simplesmente transformar de volta o intervalo de confiança e adicionar ou subtrair isso da média transformada de volta; você não pode pegar 100,344 e adicionar ou subtrair isso.

**Escolhendo a transformação certa**

As transformações de dados são uma ferramenta importante para a análise estatística adequada de dados biológicos. Para aqueles com um conhecimento limitado de estatísticas, no entanto, eles podem parecer um pouco suspeitos, uma forma de brincar com seus dados para obter a resposta desejada. Portanto, é essencial que você seja capaz de defender seu uso de transformações de dados.

Há um número infinito de transformações que você pode usar, mas é melhor usar uma transformação que outros pesquisadores costumam usar em seu campo, como a transformação de raiz quadrada para dados de contagem ou a transformação de log para dados de tamanho. Mesmo que uma transformação obscura da qual poucas pessoas tenham ouvido falar forneça dados um pouco mais normais ou mais homocedásticos, provavelmente será melhor usar uma transformação mais comum para que as pessoas não suspeitem. Lembre-se de que seus dados não precisam ser perfeitamente normais e homocedásticos; os testes paramétricos não são extremamente sensíveis a desvios de suas suposições.

Também é importante que você decida qual transformação usar antes de fazer o teste estatístico. Tentar diferentes transformações até encontrar uma que lhe dê um resultado significativo é trapaça. Se você tiver um grande número de observações, compare os efeitos de diferentes transformações na normalidade e na

homocedasticidade da variável. Se você tiver um pequeno número de observações, talvez não consiga ver muito efeito das transformações na normalidade e homocedasticidade; nesse caso, você deve usar qualquer transformação que as pessoas em seu campo usem rotineiramente para sua variável. Por exemplo, se você está estudando a distância de dispersão do pólen e outras pessoas rotineiramente a transformam em log, você também deve transformar em log a distância do pólen, mesmo que você tenha apenas 10 observações e, portanto, não possa realmente ver a normalidade com um histograma.

Se você decidir que seus dados devem seguir uma distribuição normal e precisar de transformação, existem transformações de energia simples e altamente utilizadas que veremos. Eles transformam seus dados para seguir uma distribuição normal mais de perto. No entanto, é importante observar que, ao transformar dados, você perderá informações sobre o processo de geração de dados e perderá a interpretabilidade dos valores também. Você pode considerar transformar a variável de volta em uma determinada etapa de sua análise. De um modo geral, a expressão para transformação que corresponde à geração de dados é a mais adequada. O logaritmo deve ser usado se os efeitos de geração de dados forem multiplicativos e os dados seguirem a ordem de grandezas. As raízes devem ser usadas se a geração de dados envolver efeitos quadrados.

**Transformações comuns**

Transfromação Monotônica

O que é Transformação Monotônica? É um método de transformação de dados que aplica a função matemática a cada um dos valores de dados independentemente dos outros dados. A palavra monotônica veio do procedimento do método, que transforma os valores dos dados sem alterar sua classificação. Em um termo mais simples, a transformação monotônica alterou seus dados sem depender de outros dados e não alterou sua classificação na coluna.

Um exemplo da renomada função de transformação monotônica é a transformação logarítmica ou transformação logarítmica. Assim como o nome indica, a Transformação de Log altera seu valor de dados em seus valores logarítmicos

aplicando uma função de log a cada valor de dados. Muitas variáveis seguem distribuições log-normais, o que significa que os valores seguiriam uma distribuição normal após a transformação de log. Este é um dos benefícios da Transformação de Log — seguir a suposição de normalidade, ou pelo menos próximo.

Em um termo matemático, Log Transformação é expresso na equação abaixo.

$$b_{ij} = log(x_{ij})$$

Where $b_i j$ = transformed value that replace $x_{ij}$ and $x_{ij}$ = original values

Existem muitas transformações que são usadas ocasionalmente em biologia; aqui estão três dos mais comuns:

## Transformação de log

Logaritmo log(x). Transformação comumente usada, a força dessa transformação pode ser um pouco alterada pela raiz do logaritmo. Não pode ser usado em números negativos ou 0, aqui você precisa deslocar todos os dados adicionando pelo menos |min(x)|+1.

Isso consiste em tomar o log de cada observação. Você pode usar logs de base 10 (LOG em uma planilha, LOG10 em SAS) ou logs de base e, também conhecidos como logs naturais (LN em uma planilha, LOG em SAS). Não faz diferença para um teste estatístico se você usa logs de base 10 ou logs naturais, porque eles diferem por um fator constante; o logaritmo de base 10 de um número é apenas 2,303...×0\; natural\; registro\; do\; a\; número . Você deve especificar qual log está usando ao escrever os resultados, pois isso afetará coisas como a inclinação e a interceptação em uma regressão. Eu prefiro logs de base 10, porque é possível olhar para eles e ver a magnitude do número original: log(1)=0,log(10)=1,log(100)=2 , etc.

A transformação de volta é elevar 10 ou e à potência do número; se a média dos seus dados transformados em log de base 10 for 1,43 , a média da transformação reversa será 101,43=26,9 (em uma planilha, "=10^1,43"). Se a média dos seus dados transformados em log de base e for 3,65 , a média transformada de volta será e3,65=38,5 (em uma planilha, "=EXP(3,65)". Se você tiver zeros ou números negativos,

não poderá pegue o log, você deve adicionar uma constante a cada número para torná-los positivos e diferentes de zero. Se você tiver dados de contagem e algumas das contagens forem zero, a convenção é adicionar 0,5 a cada número.

Muitas variáveis em biologia têm distribuições log-normais, o que significa que após a transformação logarítmica, os valores são normalmente distribuídos. Isso ocorre porque se você pegar um monte de fatores independentes e multiplicá-los, o produto resultante é log-normal. Por exemplo, digamos que você plantou um monte de sementes de bordo e, 10 anos depois, você vê a altura das árvores. A altura de uma árvore individual seria afetada pelo nitrogênio no solo, pela quantidade de água, quantidade de luz solar, quantidade de danos causados por insetos, etc. Ter mais nitrogênio pode tornar uma árvore 10% maior do que uma com menos nitrogênio; a quantidade certa de água pode torná-lo 30% maior do que um com muita ou pouca água; mais luz solar pode torná-lo 20% maior; menos danos causados por insetos podem torná-la 15% maior, etc. Assim, o tamanho final de uma árvore seria uma função de nitrogênio × água × luz solar × insetos e, matematicamente, esse tipo de função acaba sendo log-normal.

### Transformação de raiz quadrada

Dados distorcidos à direita (positivos):

Raiz $\sqrt[n]{x}$. Transformação mais fraca, mais forte com raiz de ordem superior. Para números negativos, cuidado especial deve ser tomado com o sinal ao transformar números negativos.

Isso consiste em tirar a raiz quadrada de cada observação. A transformação de volta é para elevar o número ao quadrado. Se você tem números negativos, não pode tirar a raiz quadrada; você deve adicionar uma constante a cada número para torná-los todos positivos.

As pessoas costumam usar a transformação de raiz quadrada quando a variável é uma contagem de algo, como colônias bacterianas por placa de Petri, células sanguíneas passando por um capilar por minuto, mutações por geração, etc.

## Transformação Arcsine

Isso consiste em tomar o arco-seno da raiz quadrada de um número. (O resultado é dado em radianos, não em graus, e pode variar de $-\pi/2$ a $\pi/2$). Os números a serem transformados em arco-seno devem estar no intervalo de 0 a 1. Isso é comumente usado para proporções, que variam de 0 a 1, como a proporção de fêmeas de mudminnows orientais que são infestadas por um parasita. Observe que esse tipo de proporção é realmente uma variável nominal, portanto, é incorreto tratá-la como uma variável de medida, quer você a transforme ou não em arco-seno. Por exemplo, seria incorreto contar o número de mudminnows que estão ou não parasitados em cada um dos vários riachos em Maryland, tratar a proporção transformada em arco-seno de fêmeas parasitadas em cada riacho como uma variável de medição e, em seguida, realizar uma regressão linear nesses dados versus profundidade do fluxo. Isso porque as proporções de riachos com menor tamanho amostral de peixes terão um desvio padrão maior do que as proporções de riachos com amostras maiores de peixes, informação que é desconsiderada ao tratar as proporções transformadas em arco como variáveis de medição. Em vez disso, você deve usar um teste projetado para variáveis nominais; neste exemplo, você deve fazer regressão logística em vez de regressão linear. Se você insistir em usar a transformação arco-seno, apesar do que acabei de dizer, a transformação reversa é o quadrado do seno do número.

## Recíproco 1/x.

Transformação mais forte, a transformação é mais forte com expoentes mais altos, por exemplo $1/x^3$. Essa transformação não deve ser feita com números negativos e números próximos de zero, portanto, os dados devem ser deslocados de maneira semelhante à transformação de log.

## Dados distorcidos à esquerda (negativos)

Refletir dados e use a transformação apropriada para inclinação à direita. Reflita cada ponto de dados subtraindo-o do valor máximo. Adicione 1 a cada ponto de dados para evitar ter um ou vários 0 em seus dados.

## Quadrado $x^2$.

Mais forte com maior potência. Não pode ser usado com valores negativos.

Exponencial $e^x$. Transformação mais forte e pode ser usado com valores negativos. Mais forte com base mais alta.

### Dados de cauda leve e pesado

Subtraia os pontos de dados da mediana e transforme. Desvios da cauda da normalidade são geralmente menos críticos do que a assimetria e podem não precisar de transformação, afinal. A subtração da mediana define seus dados para uma mediana de o. Depois disso, use uma transformação apropriada para dados distorcidos nos desvios absolutos de o em ambos os lados. Para dados de cauda pesada, use transformações para inclinação à direita para obter a mediana e para dados de cauda leve, use transformações para inclinação à esquerda para empurrar os dados para fora da mediana.

## Transformações Automáticas

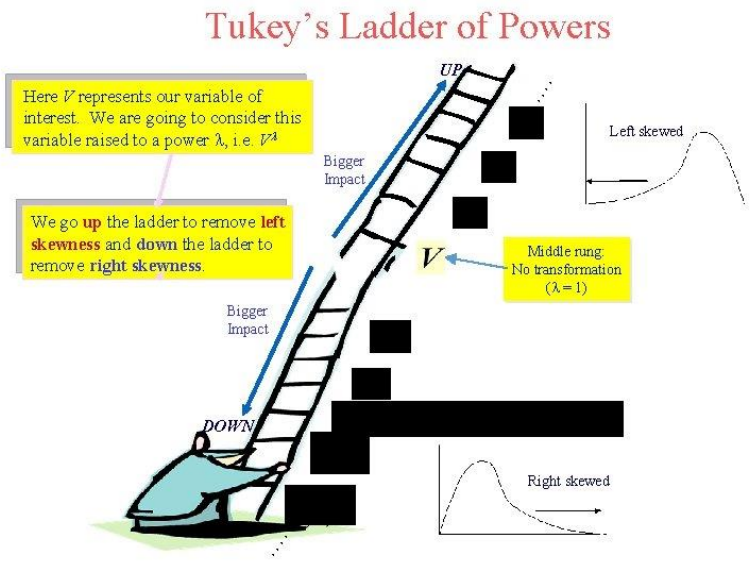
Existem várias implementações de transformações automáticas em R que escolhem a expressão de transformação ideal para você. Eles determinam um valor lambda que é o coeficiente de potência usado para transformar seus dados mais próximos de uma distribuição normal.

Use a transformada W x Gaussiana de Lambert. O pacote R LambertW tem uma implementação para transformar automaticamente dados de cauda pesada ou leve com Gaussianize().

### A Escada dos Poderes de Tukey.

Para dados distorcidos, a implementação transformTukey() do pacote R rcompanion usa testes Shapiro-Wilk iterativamente para descobrir em qual valor lambda os dados estão mais próximos da normalidade e os transforma. Os dados inclinados para a esquerda devem ser refletidos para a inclinação direita e não deve haver valores

negativos.




| $\lambda$ | -2 | -1 | -1/2 | 0 | 1/2 | 1 | 2 |
|---|---|---|---|---|---|---|---|
| y | $\frac{-1}{x^2}$ | $\frac{-1}{x}$ | $\frac{-1}{\sqrt{x}}$ | $\log x$ | $\sqrt{x}$ | $x$ | $x^2$ |

Os valores de Ladder of Powers de Tukey e as transformações de poder correspondentes. Os valores lambda podem ser decimais.

### Transformação Box-Cox.

A implementação BoxCox.lambda() da previsão do pacote R encontra iterativamente um valor lambda que maximiza a probabilidade de log de um modelo linear. No entanto, pode ser usado em uma única variável com fórmula de modelo x~1. A transformação com o valor lambda resultante pode ser feita através da função de previsão BoxCox(). Há também uma implementação no pacote R MASS. O Box-Cox padrão não pode ser usado com valores negativos, mas o Box-Cox de dois parâmetros pode.

### Transformação Yeo-Johnson.

Isso pode ser visto como uma extensão útil para o Box-Cox. É o mesmo que Box-Cox para valores não negativos e também lida com valores negativos e 0. Existem várias implementações em R via pacotes car, VGAM e receitas no framework de meta-aprendizagem de máquina arrumados.

**Relativizações (Padronização)**

Relativizações ou Padronização é um método de Transformação de Dados em que o padrão de coluna ou linha transforma os valores de dados (por exemplo, Max, Sum, Mean). É diferente da Transformação Monotônica, onde a Padronização não é independente e depende de outra estatística.

Frequentemente, você precisaria de Padronização quando ocorrer atributos com uma unidade diferente e sua análise precisa que os dados tenham uma unidade semelhante. O exemplo de análise é a análise de agrupamento ou redução de dimensionalidade, onde eles dependem da distância dos dados.

O famoso método de padronização é a padronização Z-score, onde os dados são transformados pela média e desvio padrão do recurso em escala. A média do recurso transformado seria ~0 e o desvio padrão ~1. Após a transformação de padronização do Z-score, os próprios dados transformados seriam chamados de Z-score. Em uma notação matemática, é expresso na equação abaixo.

$$z = \frac{x-\mu}{\sigma}$$

onde x = valor no recurso, μ = média do recurso e σ = desvio padrão do recurso.

Uma observação a ser lembrada, embora a padronização do Z-score tenha transformado seus dados para seguir os padrões de distribuição normal, a distribuição de recursos em si não está necessariamente seguindo a distribuição normal. Afinal, o objetivo da padronização do Z-score é redimensionar o recurso.

**Transformação probabilística (suavização)**

A transformação probabilística ou suavização é um processo de transformação de dados para eliminar quaisquer ruídos nos dados para aprimorar o padrão mais forte dentro dos dados.

A transformação é particularmente eficaz em dados heterogêneos ou ruidosos. O processo de suavização permitiu que você visse padrões de dados que antes não eram vistos. Embora você precise ter cuidado ao interpretar o resultado do processo de

suavização - ele pode mostrar que uma tendência parece confiável mesmo com dados aleatórios.

A técnica de suavização comum usada é a suavização de estimativa de densidade do kernel (KDE). Esta técnica basicamente suaviza os dados estimando a função probabilística dos dados da variável aleatória populacional com base na amostra de dados finita.

**Como transformar dados**

**Planilha**

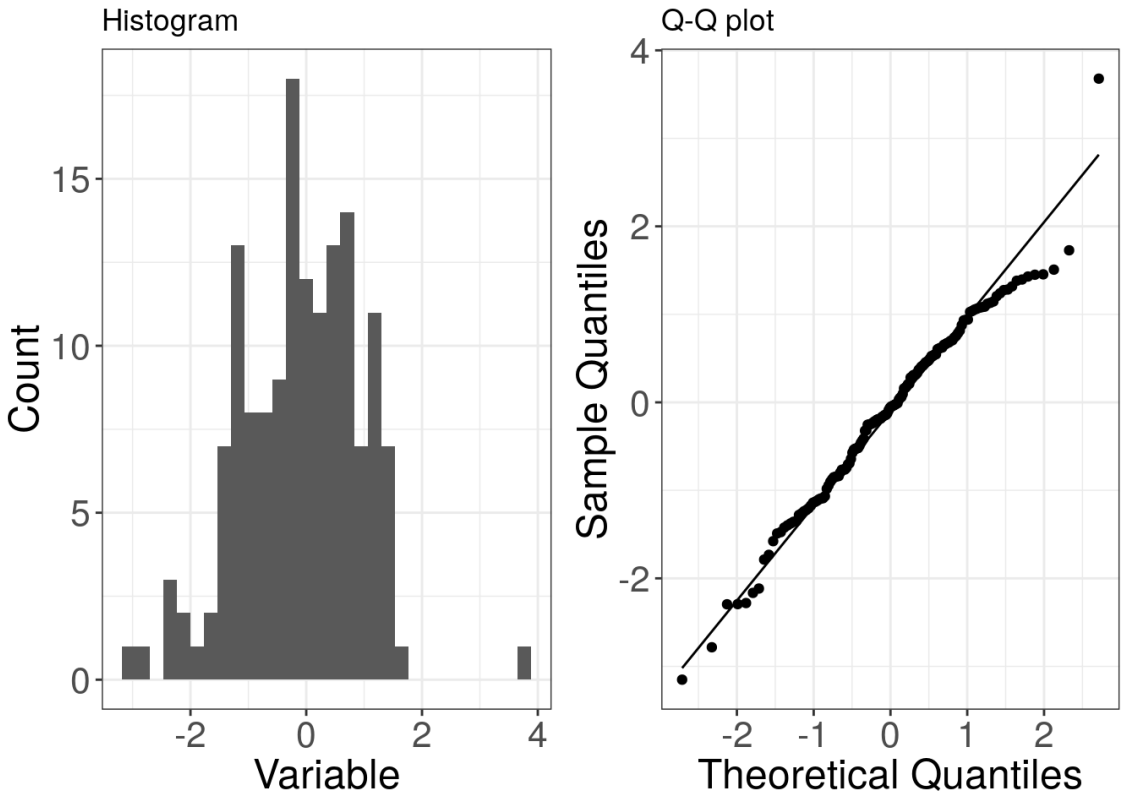
Em uma coluna em branco, insira a função apropriada para a transformação que você escolheu. Por exemplo, se você quiser transformar números que começam na célula A2 , vá para a célula B2 e digite =LOG(A2) ou =LN(A2) para transformar em log, =SQRT(A2) para transformar em raiz quadrada, ou =ASIN(SQRT(A2)) para transformação arcsine. Em seguida, copie a célula B2 e cole em todas as células da coluna B próximas às células da coluna A que contêm dados. Para copiar e colar os valores transformados em outra planilha, lembre-se de usar o comando "Colar especial..." e, em seguida, escolha colar "Valores". O uso do comando "Colar Valores Especiais...Valores" faz com que o Excel copie o resultado numérico de uma equação, em vez da equação em si. (Se sua planilha for Calc, escolha "Colar especial" no menu Editar, desmarque as caixas "Colar tudo" e "Fórmulas" e marque a caixa "Números".)

Para transformar os dados de volta, basta inserir o inverso da função que você usou para transformar os dados. Para transformar de volta os dados transformados em log na célula B2, insira =10^B2 para logs de base 10 ou =EXP(B2) para logs naturais; para dados transformados de raiz quadrada, insira =B2^2; para dados transformados em arco-seno, insira =(SIN(B2))^2

Para obter insights, os dados geralmente são transformados para seguir uma distribuição normal, seja para atender a suposições estatísticas ou para detectar relações lineares entre outras variáveis. Um dos primeiros passos para essas técnicas é verificar quão próximas as variáveis já seguem uma distribuição normal.

Como verificar se seus dados seguem uma distribuição normal?

É comum inspecionar seus dados visualmente e/ou verificar a suposição de normalidade com um teste estatístico.



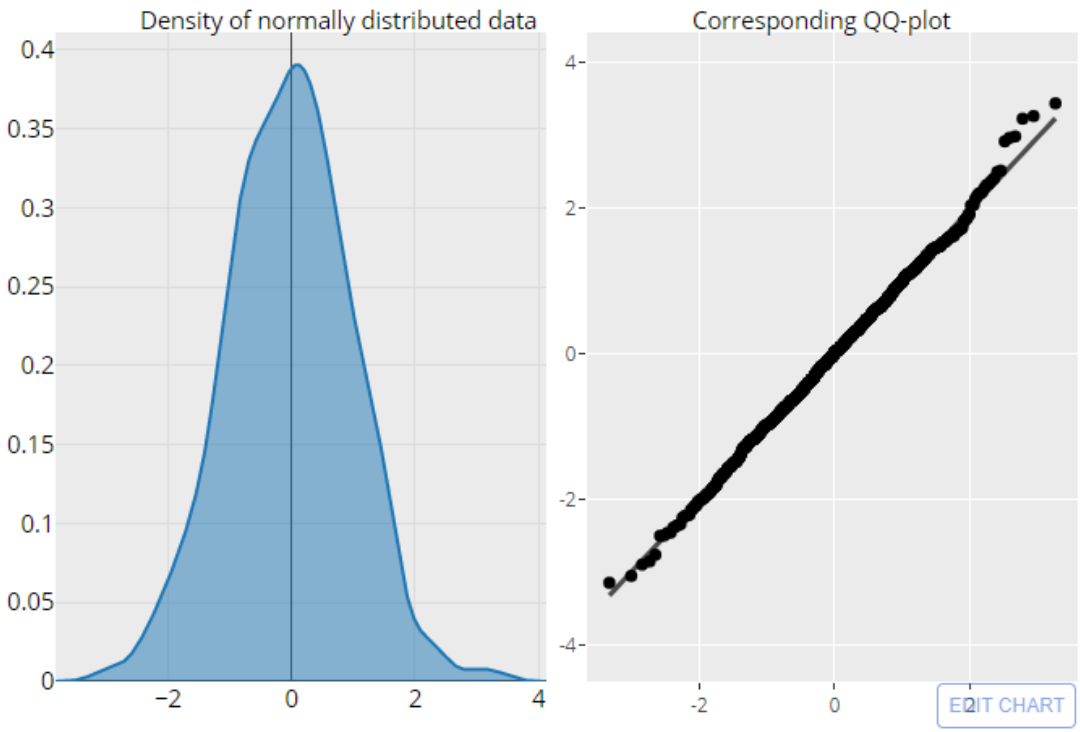

Histograma de distribuição variável e gráfico QQ correspondente com linha de referência de uma distribuição normal perfeita.

Para explorar visualmente a distribuição de seus dados, veremos o gráfico de densidade, bem como um gráfico QQ simples. O QQ-plot é uma excelente ferramenta para inspecionar várias propriedades de sua distribuição de dados e avaliar se e como você precisa transformar seus dados. Aqui, os quantis de uma distribuição normal perfeita são plotados em relação aos quantis de seus dados. Os quantis medem em que ponto de dados uma certa porcentagem dos dados é incluída. Por exemplo, o ponto de dados do quantil 0,2 é o ponto em que 20% dos dados estão abaixo e 80% estão acima. É desenhada uma linha de referência que indica como o gráfico ficaria se sua variável seguisse uma distribuição normal perfeita. Quanto mais próximos seus pontos no gráfico QQ estiverem dessa linha, mais provável será que seus dados sigam uma distribuição normal e não precisem de transformação adicional.

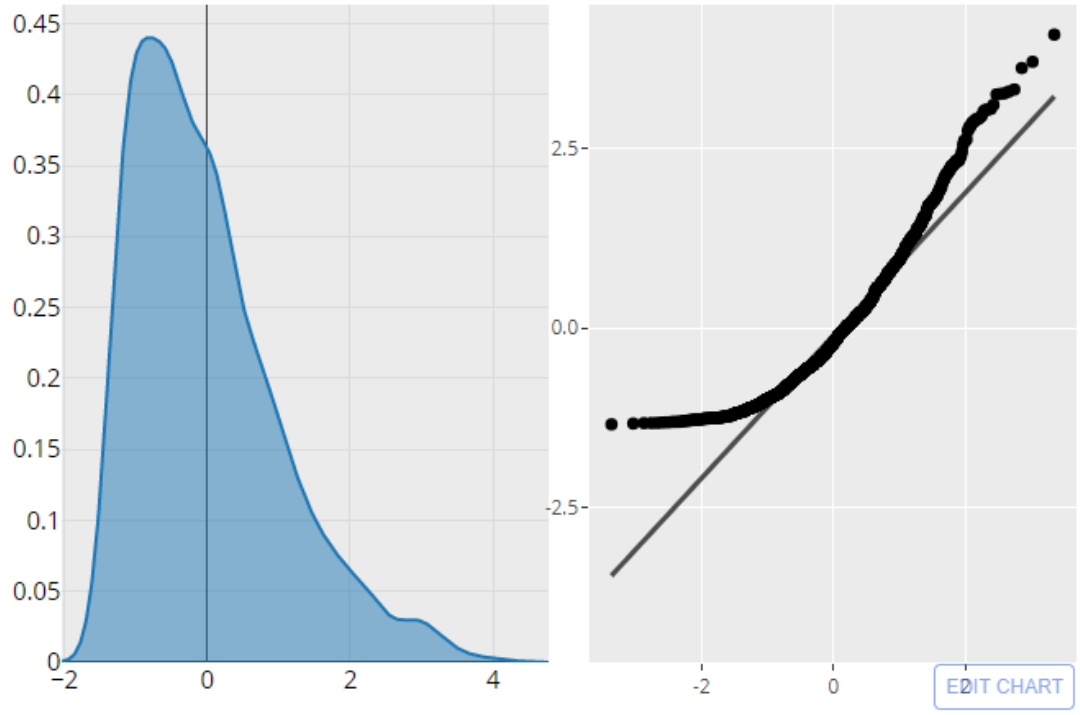
Para uma análise estatística da normalidade dos seus dados, os testes comumente usados são o Shapiro-Wilk-Test ou o Kolmogorov-Smirnov-Test. O Teste SW geralmente tem um poder de detecção maior, o Teste KS não paramétrico deve ser usado com um número elevado de observações. De um modo geral, esses testes calculam a probabilidade de sua distribuição de dados ser semelhante a uma distribuição normal (tecnicamente, qual a probabilidade de você não errar com Ho - a hipótese de que os dados são distribuídos normalmente). Esses testes, no entanto, têm os problemas bem conhecidos do Teste de Hipótese Nula Frequentista, que não está no escopo deste artigo discutir, ou seja, o problema de ser muito sensível com uma quantidade enorme de observações. O teste KS geralmente é muito sensível a pontos no meio da distribuição de dados em comparação com as caudas mais importantes. Além disso, esses testes não podem dizer o quão problemática seria uma não normalidade para obter insights de seus dados. Por isso, aconselho usar uma abordagem exploratória e visual para verificar sua distribuição de dados e renunciar a qualquer teste estatístico se você não precisar disso para um script automatizado.

### Distribuições de dados e seus gráficos QQ correspondentes

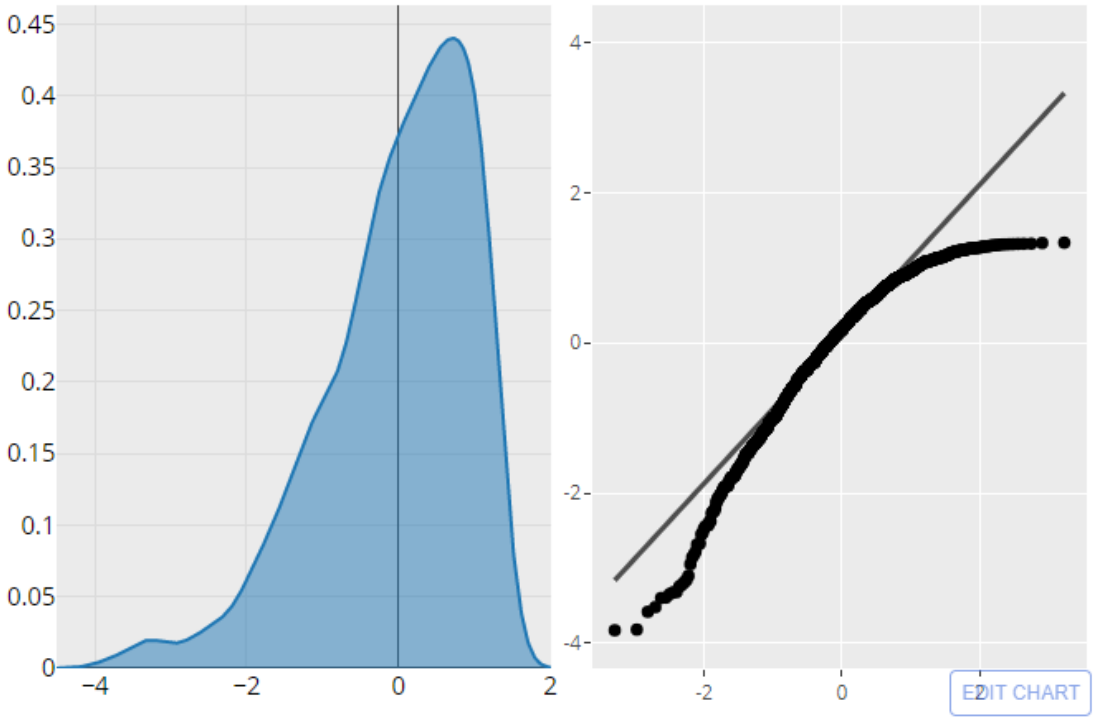
Os diagramas a seguir mostram dados simulados com a distribuição de densidade e o gráfico QQ correspondente. Quatro desvios fortes e típicos de uma distribuição normal são mostrados. Apenas para os dados normalmente distribuídos, um teste estatístico adicional para normalidade é mostrado no trecho de código para integridade.



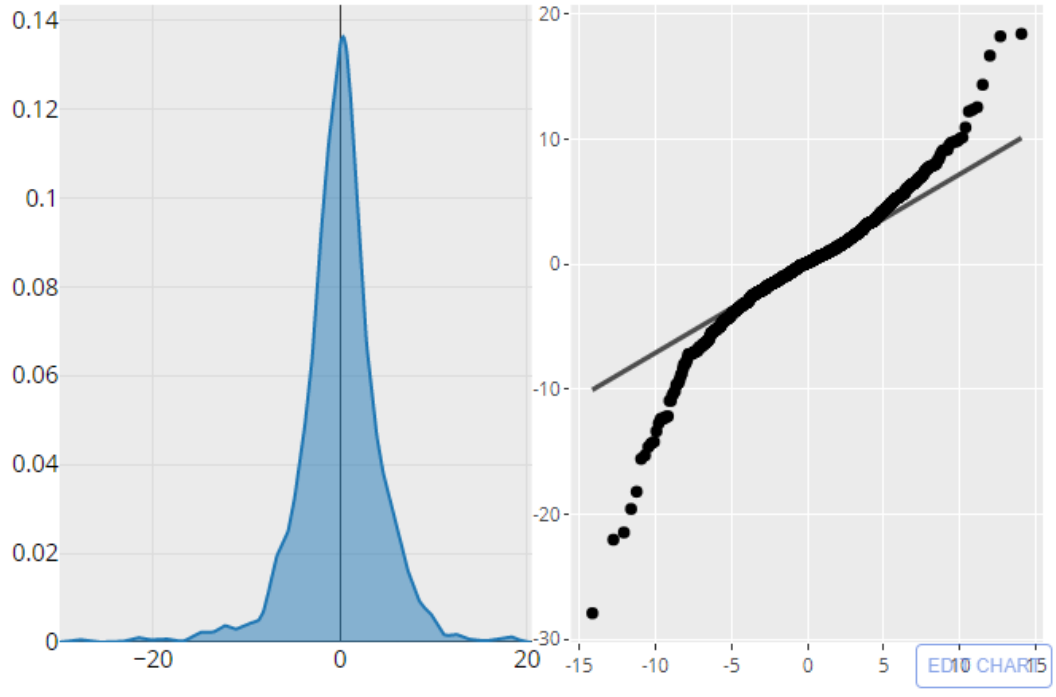

Normally distributed data and it`s QQ-plot with sample quantiles vs theoretical quantiles.



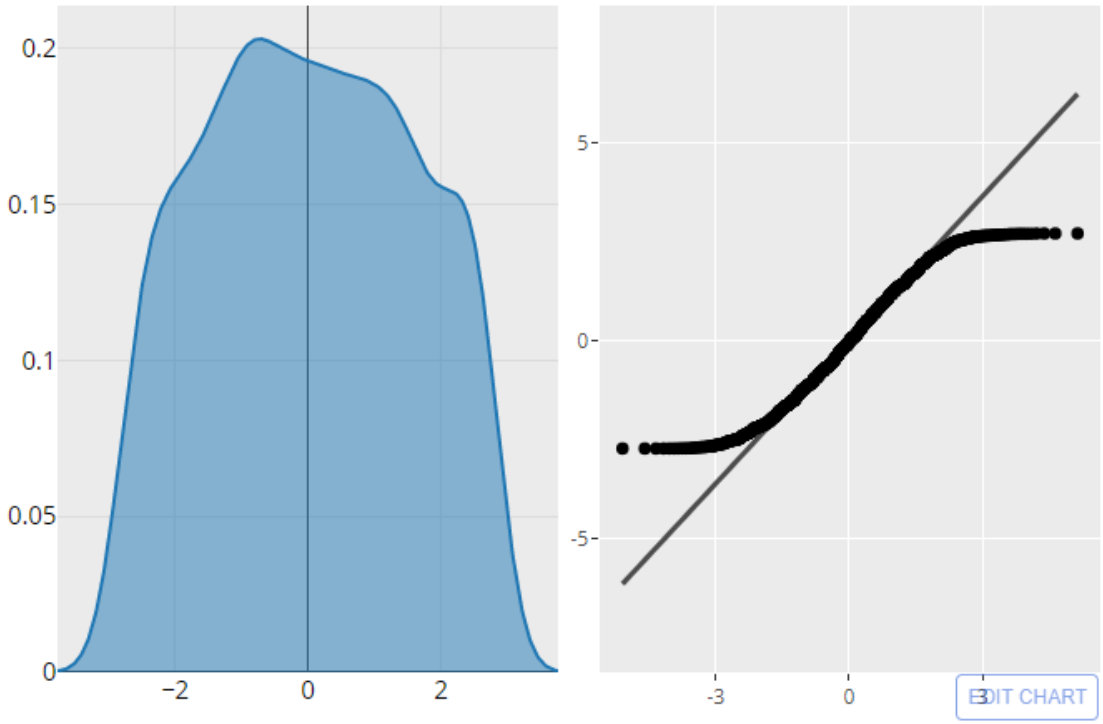

Right skewed data and corresponding QQ-plot




Left skewed data and corresponding QQ-plot




Heavy tailed (leptokurtic) data and corresponding QQ-plot




Light tailed (platykurtic) data and corresponding QQ-plot

**Comentário**

No entanto, tenha em mente que há um pouco de troca aqui. Seus dados agora podem ser normais, mas interpretar esses dados pode ser muito mais difícil. Por exemplo, se você executar um teste t para verificar as diferenças entre dois grupos e os dados que você está comparando foram transformados, você não pode simplesmente dizer que há uma diferença nas médias dos dois grupos. Agora, você tem a etapa adicional de interpretar o fato de que a diferença é baseada na raiz quadrada. Por esse motivo, geralmente tentamos evitar transformações, a menos que sejam necessárias para que a análise seja válida. Para análises como a família de testes F ou t (ou seja, testes t de amostra independente e dependente, ANOVAs, MANOVAs e regressões), as violações da normalidade geralmente não são uma sentença de morte para validade. Desde que o tamanho da amostra exceda 30 (melhor ainda se for maior que 50), geralmente não há muito impacto na validade de dados não normais; algo que Stevens enfatizou em sua publicação de 2016 da Applied Multivariate Statistics for the Social Sciences.

Esta tabela foi criada para ajudar você a decidir qual teste estatístico ou estatística descritiva é apropriado para seu experimento. Para usá-lo, você deve ser capaz de identificar todas as variáveis no conjunto de dados e dizer que tipo de variáveis elas são[6].

| test | nominal variables | measurement variables | ranked variables | purpose | notes | example |
|---|---|---|---|---|---|---|
| Exact test for goodness-of-fit | 1 | – | – | test fit of observed frequencies to expected frequencies | use for small sample sizes (less than 1000)1000) | count the number of red, pink and white flowers in a genetic cross, test fit to expected 1:2:1 ratio, total sample <1000 |
| Chi-square test of goodness-of-fit | 1 | – | – | test fit of observed frequencies to expected frequencies | use for large sample sizes (greater than 10001000) | count the number of red, pink and white flowers in a genetic cross, test fit to expected 1:2:1 ratio, total sample >1000 |
| *G*–test of goodness-of-fit | 1 | – | – | test fit of observed frequencies to expected frequencies | used for large sample sizes (greater than 1000) | count the number of red, pink and white flowers in a genetic cross, test fit to expected 1:2:1 ratio, total sample >1000 |
| Repeated *G*–tests of goodness-of-fit | 2 | – | – | test fit of observed frequencies to expected frequencies in multiple experiments | - | count the number of red, pink and white flowers in a genetic cross, test fit to expected 1:2:1 ratio, do multiple crosses |

[6]https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Biological_Statistics_(McDonald)/04%3A_Tests_for_One_Measurement_Variable/4.06%3A_Data_Transformations

| test | nominal variables | measurement variables | ranked variables | purpose | notes | example |
|---|---|---|---|---|---|---|
| Fisher's exact test | 2 | – | – | test hypothesis that proportions are the same in different groups | use for small sample sizes (less than 1000) | count the number of live and dead patients after treatment with drug or placebo, test the hypothesis that the proportion of live and dead is the same in the two treatments, total sample <1000 |
| Chi-square test of independence | 2 | – | – | test hypothesis that proportions are the same in different groups | use for large sample sizes (greater than 1000) | count the number of live and dead patients after treatment with drug or placebo, test the hypothesis that the proportion of live and dead is the same in the two treatments, total sample >1000 |
| *G*–test of independence | 2 | – | – | test hypothesis that proportions are the same in different groups | large sample sizes (greater than 1000) | count the number of live and dead patients after treatment with drug or placebo, test the hypothesis that the proportion of live and dead is the same in the two treatments, total sample >1000 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Cochran-Mantel-Haenszel test | 3 | – | – | test hypothesis that proportions are the same in repeated pairings of two groups | alternate hypothesis is a consistent direction of difference | count the number of live and dead patients after treatment with drug or placebo, test the hypothesis that the proportion of live and dead is the same in the two treatments, repeat this experiment at different hospitals |
| test | nominal variables | measurement variables | ranked variables | purpose | notes | example |
| Arithmetic mean | – | 1 | – | description of central tendency of data | - | - |
| Median | – | 1 | – | description of central tendency of data | more useful than mean for very skewed data | median height of trees in forest, if most trees are short seedlings and the mean would be skewed by a few very tall trees |
| Range | – | 1 | – | description of dispersion of data | used more in everyday life than in scientific statistics | - |
| Variance | – | 1 | – | description of dispersion of data | forms the basis of many statistical tests; in squared units, so not very understandable | - |
| Standard deviation | – | 1 | – | description of dispersion of data | in same units as original data, so | - |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | more understandable than variance | |
| Standard error of the mean | – | 1 | – | description of accuracy of an estimate of a mean | - | - |
| Confidence interval | – | 1 | – | description of accuracy of an estimate of a mean | - | - |
| test | nominal variables | measurement variables | ranked variables | purpose | notes | example |
| One-sample *t*–test | – | 1 | – | test the hypothesis that the mean value of the measurement variable equals a theoretical expectation | - | blindfold people, ask them to hold arm at 45°angle, see if mean angle is equal to 45° |
| Two-sample *t*–test | 1 | 1 | – | test the hypothesis that the mean values of the measurement variable are the same in two groups | just another name for one-way anova when there are only two groups | compare mean heavy metal content in mussels from Nova Scotia and New Jersey |
| One-way anova | 1 | 1 | – | test the hypothesis that the mean values of the measurement variable are the same in different groups | - | compare mean heavy metal content in mussels from Nova Scotia, Maine, Massachusetts, Connecticut, New York and New Jersey |
| Tukey-Kramer test | 1 | 1 | – | after a significant one-way anova, test for significant differences between all pairs of groups | - | compare mean heavy metal content in mussels from Nova Scotia vs. Maine, Nova Scotia vs. Massachusetts, Maine vs. Massachusetts, etc. |

| | | | | | | |
|---|---|---|---|---|---|---|
| Bartlett's test | 1 | 1 | – | test the hypothesis that the standard deviation of a measurement variable is the same in different groups | usually used to see whether data fit one of the assumptions of an anova | compare standard deviation of heavy metal content in mussels from Nova Scotia, Maine, Massachusetts, Connecticut, New York and New Jersey |
| test | nominal variables | measurement variables | ranked variables | purpose | notes | example |
| Nested anova | 2+ | 1 | – | test hypothesis that the mean values of the measurement variable are the same in different groups, when each group is divided into subgroups | subgroups must be arbitrary (model II) | compare mean heavy metal content in mussels from Nova Scotia, Maine, Massachusetts, Connecticut, New York and New Jersey; several mussels from each location, with several metal measurements from each mussel |
| Two-way anova | 2 | 1 | – | test the hypothesis that different groups, classified two ways, have the same means of the measurement variable | - | compare cholesterol levels in blood of male vegetarians, female vegetarians, male carnivores, and female carnivores |
| Paired *t*–test | 2 | 1 | – | test the hypothesis that the means of the continuous variable are the same in paired data | just another name for two-way anova when one nominal variable represents pairs of observations | compare the cholesterol level in blood of people before vs. after switching to a vegetarian diet |

| | | | | | | |
|---|---|---|---|---|---|---|
| Wilcoxon signed-rank test | 2 | 1 | – | test the hypothesis that the means of the measurement variable are the same in paired data | used when the differences of pairs are severely non-normal | compare the cholesterol level in blood of people before vs. after switching to a vegetarian diet, when differences are non-normal |

| test | nominal variables | measurement variables | ranked variables | purpose | notes | example |
|---|---|---|---|---|---|---|
| Linear regression | – | 2 | – | see whether variation in an independent variable causes some of the variation in a dependent variable; estimate the value of one unmeasured variable corresponding to a measured variable | - | measure chirping speed in crickets at different temperatures, test whether variation in temperature causes variation in chirping speed; or use the estimated relationship to estimate temperature from chirping speed when no thermometer is available |
| Correlation | – | 2 | – | see whether two variables covary | - | measure salt intake and fat intake in different people's diets, to see if people who eat a lot of fat also eat a lot of salt |
| Polynomial regression | – | 2 | – | test the hypothesis that an equation with $X_2X_2$, $X_3X_3$, etc. fits the YY variable significantly better than a linear regression | - | - |

| | | | | | | |
|---|---|---|---|---|---|---|
| Analysis of covariance (ancova) | 1 | 2 | – | test the hypothesis that different groups have the same regression lines | first test the homogeneity of slopes; if they are not significantly different, test the homogeneity of the YY-intercepts | measure chirping speed vs. temperature in four species of crickets, see if there is significant variation among the species in the slope or YY-intercept of the relationships |
| test | nominal variables | measurement variables | ranked variables | purpose | notes | example |
| Multiple regression | – | 3+ | – | fit an equation relating several XX variables to a single YY variable | - | measure air temperature, humidity, body mass, leg length, see how they relate to chirping speed in crickets |
| Simple logistic regression | 1 | 1 | – | fit an equation relating an independent measurement variable to the probability of a value of a dependent nominal variable | - | give different doses of a drug (the measurement variable), record who lives or dies in the next year (the nominal variable) |
| Multiple logistic regression | 1 | 2+ | – | fit an equation relating more than one independent measurement variable to the probability of a value of a dependent nominal variable | - | record height, weight, blood pressure, age of multiple people, see who lives or dies in the next year |
| test | nominal variables | measurement variables | ranked variables | purpose | notes | example |

| | | | | | | |
|---|---|---|---|---|---|---|
| Sign test | 2 | – | 1 | test randomness of direction of difference in paired data | - | compare the cholesterol level in blood of people before vs. after switching to a vegetarian diet, only record whether it is higher or lower after the switch |
| Kruskal–Wallis test | 1 | – | 1 | test the hypothesis that rankings are the same in different groups | often used as a non-parametric alternative to one-way anova | 40 ears of corn (8 from each of 5 varieties) are ranked for tastiness, and the mean rank is compared among varieties |
| Spearman rank correlation | – | – | 2 | see whether the ranks of two variables covary | often used as a non-parametric alternative to regression or correlation | 40 ears of corn are ranked for tastiness and prettiness, see whether prettier corn is also tastier |