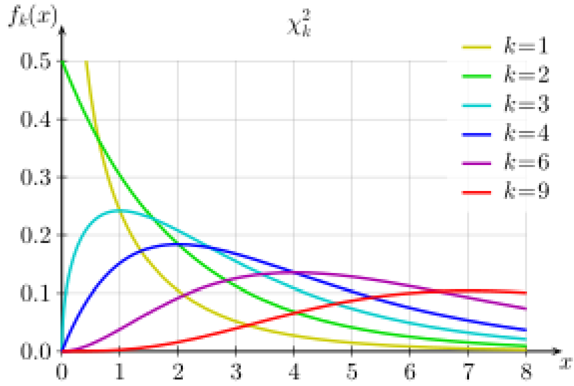
WIKIPÉDIA

Qui-quadrado

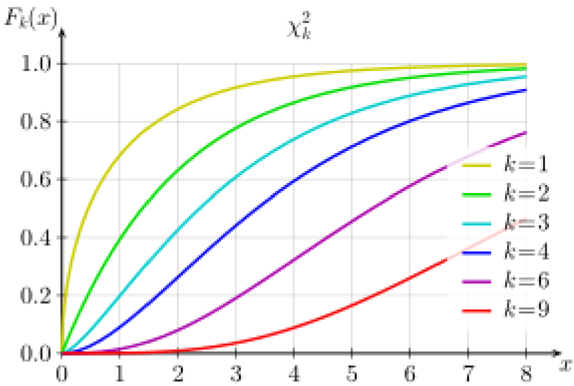
Origem: Wikipédia, a enciclopédia livre.

A distribuição χ^2 ou **qui-quadrado** é uma das distribuições mais utilizadas em estatística inferencial, principalmente para realizar **testes de χ^2**. Este teste serve para avaliar quantitativamente a relação entre o resultado de um experimento e a distribuição esperada para o fenômeno. Isto é, ele nos diz com quanta certeza os valores observados podem ser aceitos como regidos pela teoria em questão. Muitos outros testes de hipótese usam, também, a distribuição χ^2.

Distribuição Qui-quadrado



A função densidade de probabilidade da distribuição χ^2



A função distribuição acumulada da distribuição χ^2

Parâmetros	$k \in \mathbb{N}_{>0}$ graus de liberdade
Suporte	$x \in (0, +\infty)$ if $k = 1$, caso contrário $x \in [0, +\infty)$
f.d.p.	$\frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1} e^{-x/2}$
f.d.a.	$\frac{1}{\Gamma(k/2)}\, \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
Média	k
Mediana	$\approx k\left(1 - \frac{2}{9k}\right)^3$
Moda	$\max(k-2, 0)$

Índice

Teste de χ^2

Dado um experimento onde foram realizadas N medidas de uma variável aleatória X. Em cada medida, a variável X assume os valores x_1, x_2, ...,x_N. Gostaríamos de testar se a distribuição experimental dos valores x_1, x_2, ..., x_k, ..., x_N é consistente com a distribuição esperada para o fenômeno, $f(X)$. Em outras palavras, temos que avaliar como esperaríamos que as N medidas estivessem distribuídas e então comparar com a distribuição observada. Primeiramente, em geral x é uma variável contínua, de forma que não podemos nos referir ao valor esperado de medidas com um único valor de x[2] (se x for contínuo, a probabilidade de X assumir um exato valor é zero). Logo, precisamos definir intervalos $a \leq x \leq b$ e calcular o número esperado de medidas que devem estar dentro de cada

Variância	$2k$
Obliquidade	$\sqrt{8/k}$
Curtose	$\frac{12}{k}$
Entropia	$\frac{k}{2} + \log(2\Gamma(k/2)) + (1 - k/2)\psi(k/2)$
Função Geradora de Momentos	$(1 - 2t)^{-k/2}$ para $t < \frac{1}{2}$
Função Característica	$(1 - 2it)^{-k/2}$ [1]

intervalo j, em que $j = 1, 2, \ldots, n$ e n é o número de intervalos definidos. O número de medidas esperadas para o intervalo j, E_j, será, então,

$$E_j = NPr_j\,,$$

onde Pr_j é a probabilidade de X assumir um valor dentro do intervalo j. Essa probabilidade obviamente depende da distribuição $f(X)$ e é normalizada:

$$\sum_j Pr_j = 1.$$

É natural analisar a diferença entre o número de amostras observadas dentro de cada intervalo, O_j, e o número esperado:

$$O_j - E_j,$$

de forma que quanto menor forem estes valores, melhores serão as chances de nossa hipótese para $f(X)$ ser verdadeira. Porém, não podemos esperar que os dois valores O_j e E_j coincidam perfeitamente para qualquer número finito de medidas que realizarmos. Na verdade, se imaginarmos uma situação onde realizamos o procedimento de contar o número O_j muitas vezes, esperamos que a média de O_j seja E_j, com um desvio padrão $\sigma_j=E_j^{1/2}$.[2] Logo, esperamos que

$$\frac{O_j - E_j}{\sigma_j}$$

seja da ordem de unidade, se nossa hipótese for verdadeira. Definimos, portanto, a variável χ_k^2, com k graus de liberdade estatísticos, como sendo

$$\chi_k^2 \equiv \sum_{j=1}^{n} \frac{(O_j - E_j)^2}{E_j},$$

indicando o quanto as distribuições experimental e teórica são parecidas. Se $\chi^2 \leq n$, há uma boa concordância entre as distribuições, e se $\chi^2 >> n$ é bem provável que a hipótese para $f(X)$ seja falsa. Os graus de liberdade k são definidos como a diferença entre o número de medidas realizadas e o número de restrições feitas aos valores das medidas.[2]

É possível estudar as discrepâncias em experimentos que envolvam duas variáveis, em diferentes níveis. Os valores observados podem ser anotados em um quadro da seguinte forma:

Tabela das frequências de eventos com duas variáveis X e Y.

Variável X	Variável Y				Total
	Y1	**Y2**	**...**	**Yn**	
X1	O11	O12	...	O1n	L1
X2	O21	O22	...	O2n	L2
...	...	...	...	...	...
Xn	On1	On2	...	Onm	Lm
Total	C1	C2	...	Cn	T

O objetivo é observar o nível de relação existente entre as variáveis estudadas. Nesse caso, a estatística de teste é dado por:

$$\chi_k^2 \equiv \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(O_{ij}-E_{ij})^2}{E_{ij}},$$

onde as frequências esperadas são dadas por:

$$E_{ij} = \frac{L_i C_j}{T}$$

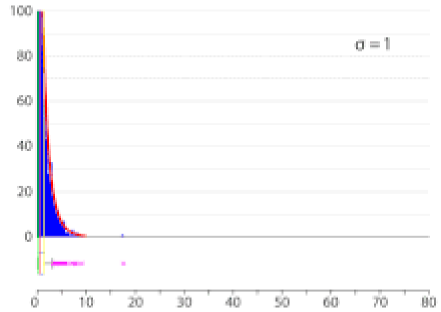
Distribuição χ^2

A probabilidade da distribuição qui quadrado não é simétrica como a da distribuição normal, para aumentar seu estado de simetria é necessário aumentar o seu grau de liberdade, portanto a relação entre simetria e grau de liberdade são diretamente proporcionais.

A variável χ_k^2, por si só, apresenta uma função densidade de probabilidade. Esta função apresenta qual a probabilidade de a variável χ_k^2 assumir um valor entre χ_k^2 e $\chi_k^2 + d\chi_k^2$, e é dada por:

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

Exemplos desta função para diversos k estão plotados na figura ao lado.



Demonstração de como a simetria cresce conforme o grau de liberdade aumenta nas distribuições qui quadrado.

Em posse desta expressão, pode-se calcular a probabilidade de, num teste de χ^2, obter-se um valor igual ou maior ao valor encontrado, $(\chi^2)'$, calculando a integral

$$\int_{(\chi^2)'}^{\infty} f(\chi^2)\, d\chi^2.$$

Desta forma, encontramos um modo quantitativo para determinar a concordância entre distribuição experimental e teórica. Em geral, para evitar o cálculo desta integral, se recorre a tabelas que apresentam os valores das probabilidades para cada intervalo de confiança e para cada grau de liberdade.

É interessante analisar que a média da distribuição χ^2 é *k*. Isto é se repetirmos o teste de χ^2 muitas vezes (para várias medidas coletadas diferentes), esperamos que a média dos valores de χ^2 encontrados tenda para o número de graus de liberdade estatísticos.

A distribuição qui-quadrado pode ser simulada a partir da distribuição normal. Por definição, se $Z_1, Z_2, \ldots Z_k$ forem k distribuições normais padronizadas (ou seja, média 0 e desvio padrão 1) independentes, então a soma de seus quadrados é uma distribuição qui-quadrado com k graus de liberdade:

$$\chi_k^2 = Z_1^2 + Z_2^2 + \ldots + Z_k^2$$

a definição é que a soma de duas qui-quadrado independentes também é uma qui-quadrado:

$$\chi_a^2 + \chi_b^2 = \chi_{a+b}^2.$$

Exemplo[*carece de fontes*?]

Podemos aplicar o teste de χ^2 para analisar quão boa é a concordância entre um conjunto de medidas (x_i, y_i) e a relação esperada $y = y(x)$.[2] Por exemplo, suponhamos que desejamos testar a hipótese de que a trajetória do lançamento de um projétil é uma parábola. O projétil sairá de uma altura de $h = 100m$, com uma velocidade inicial horizontal de $v_i = 100m/s$ e num local onde a gravidade vale $g = 9.8m/s^2$. Esperamos, portanto, que a altura do projétil em função da sua distância em relação ao ponto de partida seja:

$$y(x) = h - \frac{g}{2v_i^2}x^2.$$

Para testar a hipótese, fazemos 10 medidas de *x* e de *y* em tempos específicos. A tabela abaixo mostra os valores encontrados.

i	(x_i, y_i)
1	(50,98)
2	(80,95)
3	(110,92)
4	(140,90)
5	(170,85)
6	(200,80)
7	(230,72)
8	(260,62)
9	(290,53)
10	(320,40)

Para os valores encontrados, a incerteza na medida de *x* é desprezível e a de *y* é $\sigma = 3$. Como não calculamos nenhum parâmetro a partir dos valores medidos, o número de graus de liberdade é o mesmo do número de medidas, 10. Com estes valores, podemos calcular o valor de χ^2:

$$\chi^2_{10} = \sum_{i=1}^{10} \frac{(y_i - y(x_i))^2}{\sigma^2} = 20,24,$$

ou, ainda:

$$\frac{\chi^2_{10}}{k} = 2,024.$$

De posse do valor "normalizado" de χ^2, podemos usar uma tabela para descobrir a probabilidade de se obter este valor ou mais, e assim saber com quanta certeza podemos dizer que os valores encontradas realmente estão distribuídos como esperado. Neste caso, para 10 graus de liberdade:

$$Pr(\chi^2_{10} \geq 2,024) = 2,9\%.$$

O que descobrimos foi que a probabilidade de que as medidas obtidas realmente estejam sendo governadas pela lei prevista é de apenas 2,9%, ou seja, deveríamos rejeitar esta hipótese. Isto é, temos apenas 2,9% de certeza que a trajetória do projétil foi realmente uma parábola e que os grandes desvios observados foram apenas flutuações estatísticas.

Poderíamos ter avaliado a concordância experimental com a teórica fazendo os gráficos e comparando-os "à olho". Teríamos visto que o projétil caiu bem antes do que o previsto, sugerindo que estejamos esquecendo fatores de resistência do ar (no modelo previsto, consideramos apenas a força da gravidade, e ignoramos qualquer atrito que pudesse haver entre ar e projétil, que de fato existe, principalmente para velocidades grandes como 100 m/s).

Distribuições relacionadas

- Se *U* for uma distribuição uniforme no intervalo (0,1), então *-2 log U* é uma distribuição qui-quadrado com 2 graus de liberdade.

Nomenclatura

O símbolo χ^2, a segunda potência de χ (ou χ), envolve a forma minúscula de letra do alfabeto grego chamada *qui*[3] (também chamada, menos frequentemente, **chi** em português[4]). Devido a semelhança da letra grega com a letra xis *x* do alfabeto latino é comum a ocorrência de confusões, motivo pelo qual alguns autores optam por utilizar o nome da letra por extenso, em expressões como *qui-quadrado*. Essa é a forma recomendada pelo Glossário Inglês-Português de Estatística da Sociedade Portuguesa de Estatística e da Associação Brasileira de Estatística.[5]

Referências

1. M.A. Sanders. «Characteristic function of the central chi-squared distribution» (https://web.archive.org/web/20110715091705/http://www.planetmathematics.com/CentralChiDistr.pdf) (PDF). Consultado em 6 de março de 2009. Arquivado do original (http://www.planetmathematics.com/CentralChiDistr.pdf) (PDF) em 15 de julho de 2011
2. TAYLOR, John R., “An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements”, 1997, 2.ed.
3. «Qui» (https://www.priberam.pt/dlpo/Qui). *Dicionário Priberam da Língua Portuguesa*. Priberam Informática

4. Pedro Mateus (25 de novembro de 2010). «Sobre o nome e grafia da letra qui (alfabeto grego)» (http://www.ciberduvidas.pt/pergunta.php?id=29071). Ciberdúvidas da Língua Portuguesa. Consultado em 15 de março de 2014
5. [1] (http://glossario.spestatistica.pt/)