

Métodos Quantitativos (e Qualitativos)

5. Testes de hipóteses

Salvatore Cosentino
D.Mat. U.Minho

5 jan 2021

Testes paramétricos

Observadas umas variáveis $x, y, \dots$ um cientista pode estar interessado em **testar uma hipótese**

do tipo

$$x = y$$

(por exemplo, verificar se a velocidade da luz não depende da direção relativa às estrelas fixas),

ou

$$x > a$$

(por exemplo, decidir se o universo está em expansão, ou seja, se a constante de Hubble H é positiva ou não),

ou

x tem uma certa distribuição

(por exemplo, verificar se o decaimento radioativo é descrito por uma lei exponencial)

(por exemplo, decidir se os possíveis resultados do lançamento de um dado são equiprováveis).

Testes de comparação de grupos/amostras

Nas C.S. (ou na medicina, na biologia, ...) tipicamente estamos interessados em comparar dois ou mais grupos ou tratamentos.

Uma possibilidade é a comparação de **amostras independentes**

(um grupo recebe um tratamento e um grupo de controlo recebe um placebo, ...)

Outra possibilidade é a comparação de **amostras emparelhadas**

(a mesma amostra é avaliada antes e depois do tratamento ...)

Testes não paramétricos

Particularmente interessantes nas C.S. (ou na medicina, na biologia, ...) são testes sobre dados categóricos.

Testes de **aderência** , ou sobre a **qualidade do ajuste**

(decidir se uma distribuição de probabilidade teórica descreve bem as frequências observadas ...)

Testes de **independência**

(decidir se existem correlações entre duas variáveis, por exemplo, se ser canhoto depende ou não do sexo, ...)

Testes de **homogeneidade**

(decidir se existem diferenças na distribuição de uma variável categórica observada em duas ou mais amostras/populações ...)

Uma panóplia de testes

Existem **muitos testes**, mais ou menos sofisticados, pensado para as diferentes exigências experimentais.

Os testes mais comuns são

os **testes T de Student**
(utilizados para testes paramétricos e de comparação com variáveis numéricas, ...)

e os **testes qui-quadrado de Pearson**
(utilizados para testar aderência, independência e homogeneidade, ...)

Outros testes de interesse nas C.S. são

a **análise da variância** (em inglês, simplesmente **ANOVA**)
(para detectar diferenças significativa entre amostras/populações ...)

o **teste exato de Fisher** e o **teste de Barnard**
(para testar a independência, logo quantificar a correlação, entre duas variáveis categóricas dicotômicas)

o **teste G**
(para testar a aderência de umas frequências a umas distribuições teóricas)

Hipótese e alternativa

Como funciona um teste ?

Um teste é uma **receita estandardizada** para tomar uma **decisão**

entre uma **hipótese** H e a sua **alternativa** A

(também denotadas por H_0 , ou seja, **hipótese nula**, e H_1 , respetivamente).

dependendo dos resultados das **experiências**, ou seja, das observações de alguns observáveis.

Na prática, o cientista **conjetura** a alternativa A , e, na esperança de fazer uma descoberta científica, a compara com a hipótese H , tradicionalmente aceite pelo resto da comunidade científica, ou mais conservadora.

O modelo das observações é um modelo **probabilístico**,

e isto permite **quantificar** as diferentes probabilidades de tomar decisões corretas ou erradas dependendo se a hipótese é verdadeira ou falsa.

Estatística do teste

No modelo das observações, os resultados das experiências

$$x_1 \quad x_2 \quad x_3 \quad \dots$$

(e eventualmente outras ...) são valores observados de uma sequência de variáveis aleatórias $X_1 \quad X_2 \quad X_3 \dots$ com certas leis hipotéticas, que dependem precisamente das hipóteses que queremos testar.

O teste é tipicamente construído sobre as propriedades de apenas uma variável aleatória

$$Z = f(X_1, X_2, X_3, \dots)$$

que é uma função de todas as X_k , chamada **estatística do teste**,

e portanto a resposta do teste depende do seu valor observado

$$z = f(x_1, x_2, x_3, \dots)$$

O ponto é que os probabilistas **sabem calcular**, ou pelo menos estimar, a lei de Z no caso da hipótese H ser verdadeira, ...

... e com oportunas hipóteses razoáveis sobre os dados recolhidos.

Projetação

Naturalmente, para que as hipóteses razoáveis sobre os dados recolhidos, que justificam as receitas dos testes, sejam satisfeitas,

e para evitar falsos julgamentos, erros graves, ou até potencialmente perigosos,

é muito importante a **projetação** (em inglês, **design**) do teste, ou seja, quais dados são recolhidos e de que forma ...

Este aspeto dos testes tem muito a ver com as especificidades das diferentes áreas de pesquisa.

Por exemplo, os médicos desenvolveram **protocolos** para testar a eficácia de medicamentos, vacinas, tratamentos, ...

(cego, duplo-cego, ...)

Podem ler as **boas práticas** e os protocolos adotados nas C.S. no

Chapter 14: Analyzing Quantitative Data

de **R.D. Bachman & R.K. Schutt**, **The practice of Research in Criminology and Criminal Justice**, SAGE, 2018.

Região crítica

Fazer um teste consiste em calcular uma **região crítica** do teste (dita também **região de rejeição**), o conjunto R (tipicamente um intervalo ou uma reunião disjunta de dois intervalos) dos possíveis valores da estatística Z que consideramos não aceitável se a hipótese for verdadeira.

O complementar desta região é dita **região de aceitação** do teste.

A **receita** do teste é então a seguinte:

se $z \notin R$ **aceitamos** a hipótese H .

ou seja, aceitamos o STATUS QUO.

Vice-versa,

se $z \in R$ **rejeitamos** a hipótese H

logo **aceitamos a alternativa** A , e publicamos a nossa descoberta numa revista!

O cálculo da região crítica deve ser feito de uma forma **estandardizada**, ou seja, codificada e aceite pela comunidade científica da área.

Como calcular a região crítica ?

Nível de significância

O **nível de significância** α do teste é a maior das probabilidades

$$\alpha = \max \mathbf{P}(\text{rejeitar } H \mid H \text{ é verdadeira})$$

de rejeitar a hipótese se a hipótese for verdadeira.

Um cientista **honesto** testa a hipótese mais **conservadora**, ou seja, o STATUS QUO

(se quero anunciar ao mundo que a água tem memória, testo a hipótese de que a água não tem memória)

(se quero provar que um fármaco é eficaz no tratamento do Covid-19, testo a hipótese de que o fármaco tem o mesmo efeito de um placebo)

portanto rejeitar H quando H é verdadeira, chamado **erro de tipo I**, é considerado um erro **grave**.

Consequentemente, é importante utilizar valores **pequenos** do nível de significância α , tipicamente

5% ou 1%

e escolher α , que determina a região crítica R , **antes** de fazer as experiências.

Potência

O outro possível erro é aceitar a hipótese H quando a alternativa A é verdadeira, ou seja, não reconhecer uma descoberta científica.

Este é chamado **erro de tipo II**, e é considerado menos grave do erro de tipo I.

Se a sua probabilidade máxima é

$$\beta = \max \mathbf{P}(\text{aceitar } H \mid A \text{ é verdadeira})$$

então

$$1 - \beta$$

é chamado **potência do teste**.

Seria desejável ter β pequeno, e portanto potência grande (ou seja, próxima de um).

Lamentavelmente, quando β cresce então α decresce, e vice-versa!

É **prática comum** escolher a região crítica usando apenas um nível de significância α pequeno (5% ou 1%), assim sacrificando a potência $1 - \beta$.

p -value

Os software de estatística calculam diretamente o **valor de prova** (em inglês, *p-value*) a partir dos dados das experiências e portanto do valor observado z da estatística do teste (e do número de observações, ...).

O **p -value** é a probabilidade p

$$p = \mathbf{P}(\text{observar } Z \text{ pior que } z \mid H \text{ é verdadeira})$$

de observar um valor pior, ou seja, **mais extremo**, do valor atualmente observado z se a hipótese for verdadeira.

Equivalentemente, é o **menor** nível de significância com que se rejeitaria a hipótese H numa experiência onde se observou o valor z da estatística.

Isto significa que se, por exemplo, $p = 0.05$, então as nossa experiências permitem aceitar a hipótese com qualquer nível de significância $\alpha \geq 5\%$.

Assim, um p -value suficientemente **pequeno**, por exemplo

$$p < 0.01$$

é indício de que muito provavelmente é oportuno **rejeitar a hipótese** H , logo aceitar a alternativa A (e portanto anunciar uma descoberta científica!).

Exemplo: a moeda “honesta” ...

Como decidir se uma moeda é **honesta** ou **enfeitiçada** ?

O senso comum, logo a hipótese conservadora H , diz que se uma moeda é honesta então a probabilidade de sair CARA em cada lançamento é $p = 1/2$. Portanto a alternativa A , a moeda enfeitiçada, é uma moeda tal que $p \neq 1/2$.

Se eu lançar um número grande n de vezes uma moeda honesta, não é naturalmente razoável esperar um número de CARAs exatamente igual a $n/2$.

A variável S_n , que conta o número de CARAs obtidas em n lançamentos (ou também a posição de um passeio aleatório simétrico) tem **flutuações típicas** da ordem de

$$\left| S_n - \frac{n}{2} \right| \sim \sqrt{n}$$

em torno do seu valor médio $n/2$.

Mais precisamente, a teoria das probabilidades permite **quantificar** a probabilidade destas flutuações.

Por exemplo, o **teorema limite central** diz que com probabilidade pelo menos 95% esta variável está num intervalo

$$S_n \simeq \frac{n}{2} \pm \sqrt{n}$$

... ou enfeitiçada ?

Ontem a minha filha lançou $n = 100$ vezes uma moeda de um euro, obtendo um número de CARAs igual a

$$S_n = 44$$

A diferença entre o valor esperado 50 e o valor observado 44 é um efeito natural do acaso, é uma flutuação típica, pois

$$|44 - 50| \sim \sqrt{100}$$

Em outras palavras, o valor observado S_n está bem dentro do intervalo de confiança de nível 95%

$$S_n \simeq 50 \pm 10$$

Assim, o resultado de um teste da hipótese H com **nível de significância** $\alpha = 5\%$ é que a hipótese é aceitável.

A probabilidade de obter valores iguais ou ainda **mais extremos** de 53, chamada p -value, também pode ser calculada, usando a aproximação do teorema limite central. O seu valor é

$$p = \mathbf{P}(|S_n - 50| \geq 4) \simeq 0.52$$

que é uma probabilidade bastante grande.

Conclusão: não temos razões para suspeitar que a minha moeda seja enfeitiçada.

Testes sobre médias ingênuos

Uma maneira (aparentemente) ingênua de testar hipóteses sobre médias consiste em calcular **intervalos de confiança** apropriados.

Seja m o **valor verdadeiro**, ou seja, o valor esperado $\mathbb{E}X = m$ do observável x .

Para testar a hipótese ($m = m_0$) contra a **alternativa simétrica** ($m \neq m_0$), podemos calcular un intervalo de confiança centrado

$$x = \bar{x} \pm t_{1-\alpha/2} \frac{S_x}{\sqrt{n}}$$

de nível $1 - \alpha = 95\%$, e aceitar a hipótese se m_0 está neste intervalo.

Para testar a hipótese ($m \leq m_0$) contra a **alternativa assimétrica** ($m > m_0$) (natural quando $m < m_0$ não faz sentido) podemos calcular un intervalo de confiança assimétrico

$$x < \bar{x} - t_{1-\alpha} \frac{S_x}{\sqrt{n}}$$

de nível $1 - \alpha = 95\%$, e aceitar a hipótese se m_0 está neste intervalo.

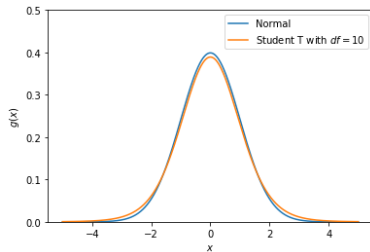
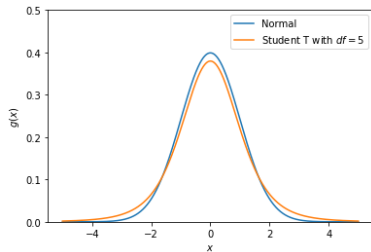
Estatística do teste T de Student

Se a lei das X_k é normal, com média m (o valor verdadeiro do observável x) e desvio padrão σ (estimado por S_x), então a variável

$$Z = \frac{\bar{X} - m}{S_x / \sqrt{n}}$$

tem **lei de Student**¹ T_{n-1} com $df = n - 1$ graus de liberdade (em inglês, **degrees of freedom**),

que é **muito bem aproximada** por uma lei normal reduzida $N(0, 1)$ quando n é grande (basta $df \geq 5$, como podem ver nas imagens).



¹pseudônimo do estatístico e químico inglês **Gosset** que, trabalhando na cervejaria Guinness em Dublin, não podia publicar artigos usando o seu próprio nome

Cálculo das probabilidades

Obtidos os resultados $x_1, x_2, \dots, x_n$ das experiências, e se a hipótese é ($m = m_0$), podemos calcular

$$z = \frac{\bar{x} - m_0}{S_x / \sqrt{n}}$$

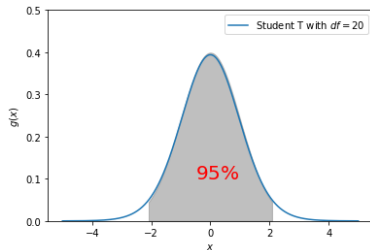
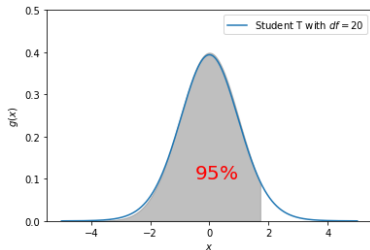
que é o **valor observado da estatística Z** do teste.

Fixado um nível de significância, por exemplo $\alpha = 5\%$, as tabelas/software calculam o valor t_α tal que

$$\mathbf{P}(Z > t_{1-\alpha}) = \alpha$$

ou o valor $t_{1-\alpha/2}$ tal que

$$\mathbf{P}(|Z| > t_{1-\alpha/2}) = \alpha$$



Receitas do teste T de Student

... e finalmente:

uma região crítica para testar a hipótese ($m = m_0$) contra a **alternativa simétrica** ($m \neq m_0$) é

$$|\bar{x} - m_0| > t_{1-\alpha/2} \frac{S_x}{\sqrt{n}}$$

uma região crítica para testar a hipótese ($m \leq m_0$) contra a **alternativa assimétrica** ($m > m_0$) é

$$\bar{x} > m_0 + t_{1-\alpha} \frac{S_x}{\sqrt{n}}$$

Testes T de Student sobre probabilidades

Um caso particular de média é uma **probabilidade** p , a probabilidade de sucesso, logo a média, em uma prova de Bernoulli.

Neste caso as observações x_k têm valores 0 ou 1, e a probabilidade $p = \mathbf{P}(x_k = 1)$ é estimada com $\bar{x}$, a frequência amostral.

Na **hipótese** ($p = p_0$), então a variável $n\bar{x}$ tem lei binomial $B(n, p)$, com esperança np_0 e variância $np_0(1 - p_0)$, logo a variável

$$Z = \frac{\bar{x} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim N(0, 1)$$

tem lei aproximadamente normal $N(0, 1)$ quando n é grande.

Consequentemente, fixado um nível de significância $\alpha = 5\%$ (ou menor), uma região crítica para testar a hipótese ($p = p_0$) contra a **alternativa assimétrica** ($p > p_0$) é

$$\bar{x} > p_0 + \phi_{1-\alpha} \frac{\sqrt{p_0(1 - p_0)}}{\sqrt{n}}$$

onde $\phi_{1-\alpha} \simeq 1.96$ é o quantil da lei normal.

Testes de comparação

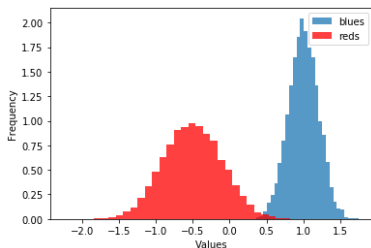
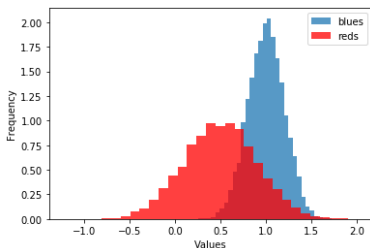
O problema é decidir se dois observáveis quantitativos, x e y , por exemplo medidos em duas amostras independentes, ou apenas numa amostra antes e depois de um tratamento, são **iguais**.

(por exemplo, decidir se a velocidade da luz na direção do movimento terrestre é igual a velocidade da luz na direção ortogonal)

Uma maneira (aparentemente) ingênua de testar a **hipótese** ($x = y$) contra a **alternativa** ($x \neq y$) consiste em calcular os dois intervalos de confiança

$$\bar{x} \pm t_{1-\alpha/2} \frac{S_x}{\sqrt{n}} \quad \text{e} \quad \bar{y} \pm t_{1-\alpha/2} \frac{S_y}{\sqrt{m}}$$

de nível suficientemente grande ($\alpha \geq 95\%$ ou 99%), e aceitar a hipótese se estes intervalos não forem claramente disjuntos.



Estatística do teste de comparação de Student

Sejam

$$x_1 \quad x_2 \quad x_3 \quad \dots \quad x_n \quad \text{e} \quad y_1 \quad y_2 \quad y_3 \quad \dots \quad y_m$$

as observações.

Se as leis dos x_k e dos y_k são normais, com variâncias σ_x^2 e σ_y^2 , respectivamente, então a variável

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

tem lei normal $N(0, 1)$ na **hipótese** ($x = y$).

As variâncias podem ser estimadas com as variâncias amostrais S_x^2 e S_y^2 .

Se

$$S_{\text{tot}}^2 = \frac{1}{n + m - 2} ((n - 1)S_x^2 + (m - 1)S_y^2)$$

denota a **variância total**, então a variável

$$z = \frac{\bar{x} - \bar{y}}{S_{\text{tot}} \sqrt{1/n + 1/m}}$$

tem lei de Student T_{n+m-2} com $df = n + m - 2$ graus de liberdade.

Receita do teste de comparação Student

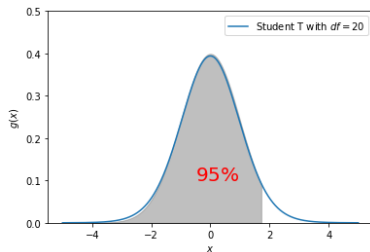
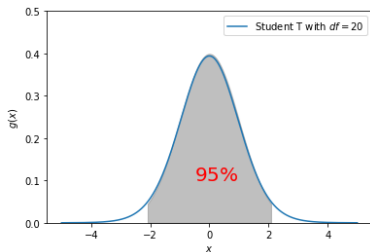
Finalmente, fixado um nível de confiança, por exemplo $\alpha = 5\%$,

uma região crítica para testar a hipótese ($x = y$) contra a **alternativa simétrica** ($x \neq y$) é

$$|\bar{x} - \bar{y}| > t_{1-\alpha/2} S_{\text{tot}} \sqrt{\frac{1}{n} + \frac{1}{m}}$$

e uma região crítica para testar a hipótese ($x \leq y$) contra a **alternativa assimétrica** ($x > y$) é

$$\bar{x} - \bar{y} > t_{1-\alpha} S_{\text{tot}} \sqrt{\frac{1}{n} + \frac{1}{m}}$$



Testes de aderência

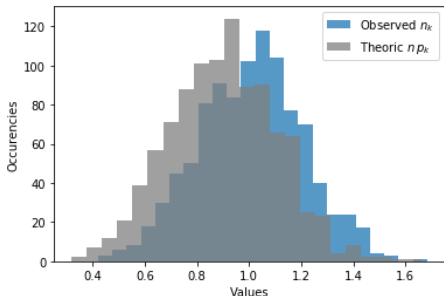
O problema é testar um **modelo probabilístico**.

(por exemplo, decidir se um dado é “honesto”, ou seja, mostra as suas faces com probabilidades iguais ...)

A **hipótese** H é: o observável X assume os valores $x_1, x_2, \dots, x_M$ (que podem ser classes) com probabilidades

$$p_k = \mathbf{P}(X = x_k)$$

Uma ideia natural é comparar os histogramas das observações com o histograma teórico, ou seja, com a função massa de probabilidades.



Flutuações esperadas

Sejam $f_1, f_2, \dots, f_M$ as **frequências empíricas** em n observações, ou seja

$$f_k = \frac{n_k}{n}$$

onde n_k o número de vezes que observamos o valor/classe x_k .

A **lei dos grandes números** sugere que

$$f_k \sim p_k$$

logo $n_k \sim np_k$, quando n é grande.

O **teorema limite central** sugere que os **desvios quadráticos observados**

$$(n_k - np_k)^2$$

sejam da ordem (ou seja, uma ou duas . . . vezes) das **variâncias esperadas**

$$np_k(1 - p_k) \simeq np_k$$

(se os p_k 's são pequenos).

Qui-quadrado esperado

Uma medida global das **flutuações observadas** é a soma dos quocientes

$$Q^2 = \sum_{k=1}^M \frac{(n_k - np_k)^2}{np_k}$$

Valores **esperados** de Q^2 no caso da hipótese ser verdadeira são portanto da ordem de

$$Q^2 \sim M$$

Um **teorema de Pearson** diz que, quando n é grande, as variáveis Q^2 têm leis bem aproximadas pela lei de uma variável qui-quadrado $\chi^2(M-1)$ com $df = M-1$ graus de liberdade.

Os estatísticos concordam em dizer que a aproximação de Pearson começa a ser boa (e portanto o teste é significativo) desde que os números esperados np_k de observações em cada uma das classes sejam

$$np_k \geq 5$$

Se o número de graus de liberdade é mesmo grande, como $df \geq 50$, então uma variável qui-quadrado é, com probabilidade 95%, limitada por

$$Q^2 \leq M + 2\sqrt{2M}$$

Receita do teste qui-quadrado sobre a aderência

Uma região crítica do teste de aderência com nível de significância α é portanto

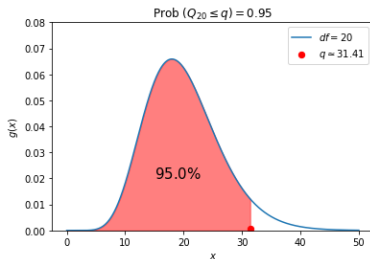
$$Q^2 > q_{1-\alpha}$$

onde $q_{1-\alpha}$ é o quantil de uma variável $\chi^2(nM1)$.

Em alternativa, um software calcula diretamente o p -value, ou seja, a probabilidade p de obter um qui-quadrado superior ao valor observado Q^2 . Então

se $p < \alpha$ rejeitamos a hipótese H

logo concluímos que a lei conjecturada não descreve bem os resultados da experiência.



Testes de independência

O problema é testar a **independência** entre dois (ou mais) observáveis, por exemplo, entre duas características de uma população.

(por exemplo, decidir se, ao lançar um dado e uma moeda, o número de pintas que mostra o dado é independente da face que mostra a moeda ...)

(por exemplo, decidir se a probabilidade de desenvolver uma doença depende ou não da presença de um determinado gene ...)

Na linguagem das C.S., o objetivo é procurar alguma forma de **associação** entre duas características de uma população

(por exemplo, determinar os fatores que influenciam o desenvolvimento de um comportamento criminoso ...)

Naturalmente, associação não significa necessariamente **dependência** !

A dicotomia correta é entre **independência** e **correlação**, objeto da próxima e última aula.

Independência

Apesar de ser uma palavra do dia a dia, e ter até festividades dedicadas, para matemáticos e físicos a **independência** tem um significado preciso e quantificável (que é de fato a ideia central da teoria das probabilidades).

O que significa, por exemplo, dizer que “ao lançar um dado e uma moeda (que nos achamos não falarem entre si), o número de pintas que mostra o dado é independente da face que mostra a moeda” ?

Dizer que a probabilidade de sair, por exemplo, UMA PINTA é $p = 1/6$ significa que o dado mostra UMA PINTA aproximadamente uma vez em cada 6 lançamentos.

Por outro lado, dizer que a probabilidade de sair CARA é $q = 1/2$ significa que a moeda mostra CARA aproximadamente uma vez em cada 2 lançamentos.

Assim, se lançamos 12 vezes um dado e uma moeda, esperamos observar $12 \cdot p = 2$ vezes UMA PINTA e, destas duas vezes, apenas $2 \cdot q = 1$ vez CARA.

Assim, a probabilidade de observar contemporaneamente CARA e UMA PINTA é

$$\frac{1}{12} = p \cdot q$$

Em geral, as duas variáveis X e Y são **independentes** se

$$\mathbf{P}(X = x_i \text{ e } Y = y_j) = \mathbf{P}(X = x_i) \cdot \mathbf{P}(Y = y_j)$$

Tabelas de contingência

Se as variáveis X e Y assumem valores/classes

$$x_1, \dots, x_M \quad \text{e} \quad y_1, \dots, y_N$$

respetivamente, então o resultado de uma experiência é uma **tabela de contingência**, que conta os números

$$n_{ij} = \#\{\text{observações tais que } X = x_i \text{ e } Y = y_j\}$$

num total de n observações.

	y_1	y_2	y_3	$\dots$	total
x_1	n_{11}	n_{12}	n_{13}		n_{1*}
x_2	n_{21}	n_{22}	n_{23}		n_{2*}
x_3	n_{31}	n_{32}	n_{33}		n_{3*}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		
total	n_{*1}	n_{*2}	n_{*3}		n

Também, podemos calcular os números de vezes

$$n_{i*} = \sum_{j=1}^N n_{ij} \quad \text{e} \quad n_{*j} = \sum_{i=1}^M n_{ij}$$

em que foram observados os eventos $(X = x_i)$ ou $(Y = y_j)$, respetivamente, assim que o número total das observações é

$$n = \sum_i n_{i*} = \sum_j n_{*j}$$

Frequências observadas e esperadas

As probabilidades $\mathbf{P}(X = x_i)$ e $\mathbf{P}(Y = y_j)$ são estimadas pelas frequências observadas

$$f_{i*} = \frac{n_{i*}}{n} \quad \text{e} \quad f_{*j} = \frac{n_{*j}}{n}$$

dos eventos $(X = x_i)$ ou $(Y = y_j)$, respetivamente.

e as probabilidades $\mathbf{P}(X = x_i \text{ e } Y = y_j)$ são estimadas pelas frequências observadas

$$f_{ij} = \frac{n_{ij}}{n}$$

Na hipótese de independência, é natural esperar que

$$f_{ij} \sim f_{i*} f_{*j}$$

e portanto que as ocorrências esperadas dos eventos $(X = x_i \text{ e } Y = y_j)$ sejam da ordem de

$$n f_{i*} f_{*j}$$

Esta expectativa deve ser comparada com as ocorrências observadas n_{ij} .

Flutuações e qui-quadrado esperado

O **teorema limite central** sugere que as **flutuações** das ocorrências observadas n_{ij} em torno do valor esperado $n f_{i*} f_{*j}$,

$$(n_{ij} - n f_{i*} f_{*j})^2$$

sejam da ordem das variâncias esperadas

$$n f_{i*} f_{*j} (1 - f_{i*} f_{*j}) \simeq n f_{i*} f_{*j}$$

Consequentemente, esperamos que a soma dos quocientes

$$Q^2 = \sum_{ij} \frac{(n_{ij} - n f_{i*} f_{*j})^2}{n f_{i*} f_{*j}}$$

a **estatística do teste qui-quadrado**, seja da ordem de

$$Q^2 \sim MN$$

o número total de classes das duas variáveis (ou seja, o número de entradas da tabela de contingência), e não muito maior.

Um **teorema de Pearson** diz que, quando n é grande, a variável Q^2 tem lei bem aproximada pela lei de uma variável **qui-quadrado** $\chi^2((M-1)(N-1))$, com $df = (M-1)(N-1)$ graus de liberdade.

Receita do teste qui-quadrado para a independência

Uma região crítica de um teste de independência com nível de significância α é portanto

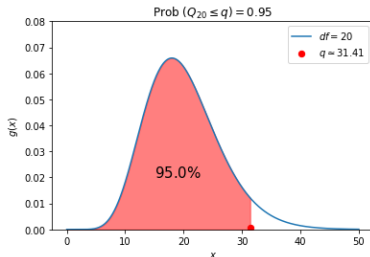
$$Q^2 > q_{1-\alpha}$$

onde $q_{1-\alpha}$ é o quantil da lei qui-quadrado.

Em alternativa, um software calcula o p -value, ou seja, a probabilidade p de obter um qui-quadrado superior ao valor observado Q^2 . Então

se $p < \alpha$ rejeitamos a hipótese H

e portanto concluímos que as variáveis não são independentes.



Coefficiente Φ

Quando as variáveis são **dicotómica**, logo a tabela de contingência é

	$y = 0$	$y = 1$	total
$x = 0$	n_{00}	n_{01}	n_{0*}
$x = 1$	n_{10}	n_{11}	n_{1*}
total	n_{*0}	n_{*1}	n

uma medida da falta de independência, ou seja, da **correlação** (que nas C.S. é chamada **associação**) entre as variáveis é o **coeficiente Φ**

$$\Phi = \frac{n_{00} n_{11} - n_{01} n_{10}}{\sqrt{n_{0*} n_{1*} n_{*0} n_{*1}}}$$

(conhecido como **Matthews Correlation Coefficient** nas áreas da bioquímica e do **machine learning**), que é um caso particular do **coeficiente de correlação de Pearson**², e assume valores entre -1 e 1 .

O seu quadrado é

$$\Phi^2 = \frac{Q^2}{n}$$

Valores $\Phi \simeq 0$ indicam independência, e valores $\Phi \simeq \pm 1$ indicam correlação, positiva ou negativa, respetivamente.

²tratado na próxima aula

Exemplo: um dado e uma moeda ...

A minha filha lançou $n = 120$ vezes um dado e uma moeda, obtendo os resultados

	UM	DOIS	TRÊS	QUATRO	CINCO	SEIS	total
CARA	10	9	11	12	6	5	53
COROA	8	16	13	7	7	16	67
total	18	25	24	19	13	21	120

Com probabilidade 95%, uma **moeda honesta** mostra CARA um número de vezes

$$n \frac{1}{2} \pm 2 \sqrt{n \frac{1}{2} \left(1 - \frac{1}{2}\right)} \simeq 60 \pm 11$$

um **dado honesto** mostra cada uma das suas faces um número de vezes

$$n \frac{1}{6} \pm 2 \sqrt{n \frac{1}{6} \left(1 - \frac{1}{6}\right)} \simeq 20 \pm 8$$

e um dado e uma moeda **honestos e independentes** mostram cada um dos 12 resultados possíveis um número de vezes

$$n \frac{1}{12} \pm 2 \sqrt{n \frac{1}{12} \left(1 - \frac{1}{12}\right)} \simeq 10 \pm 6$$

Os resultados da experiência são claramente **compatíveis** com estes intervalos.

... honestos e independentes

Os testes qui-quadrado não podem que **confirmar** as nossas observações.

Um teste de aderência sobre a moeda, a hipótese nula sendo **equiprobabilidade** entre CARA e COROA, tem como resultados

$$Q^2 \simeq 1.63 \quad \text{e} \quad p \simeq 0.20$$

Um teste de aderência sobre a dado, a hipótese nula sendo **equiprobabilidade** entre UMA, DUAS, TRÊS, QUATRO, CINCO e SEIS PINTAS, tem como resultados

$$Q^2 \simeq 4.80 \quad \text{e} \quad p \simeq 0.44$$

Finalmente, um teste de **independência** sobre dado e moeda tem como resultado

$$Q^2 \simeq 7.98 \quad \text{e} \quad p \simeq 0.16$$

Testes de homogeneidade

O problema é testar a **homogeneidade** entre duas ou mais populações em relação a uma certa variável, que mede uma certa **caraterística** dos indivíduos.

(por exemplo, decidir se a distribuição dos ordenados pro capite é igual nos diferentes países da UE ...)

(por exemplo, decidir se uma terapia tem os mesmos efeitos em pacientes com diferentes tipos de uma determinada doença ...)

É essencialmente um teste de independência entre duas variáveis X e Y se consideramos que a variável X determina a população, ou seja, que os seus valores

$$x_1 \quad x_2 \quad \dots \quad x_M$$

são apenas os **nomes/classificadores** das diferentes populações, e que portanto podemos substituir por $1, 2, 3, \dots, M$.

Naturalmente, de cada população é escolhida uma **amostra**, de tamanho razoável e de forma possivelmente aleatória, de acordo com os protocolos usuais das técnicas de amostragem.

Tabelas de contingência

Se a variável Y assume os valores/classes

$$y_1 \quad y_2 \quad \dots \quad y_N$$

respetivamente, então o resultado de uma experiência é uma **tabela de contingência**, que conta os números

$$n_{ij} = \#\{\text{observações de } Y = y_j \text{ dentro da } i\text{-ésima população}\}$$

	y_1	y_2	y_3	$\dots$	total
1	n_{11}	n_{12}	n_{13}		n_1
2	n_{21}	n_{22}	n_{23}		n_2
3	n_{31}	n_{32}	n_{33}		n_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$		
M	n_{M1}	n_{M2}	n_{M3}		n_M
total	n_{*1}	n_{*2}	n_{*3}		n

Neste caso as somas

$$n_i = \sum_{j=1}^N n_{ij}$$

são os **tamanhos** das diferentes amostras das M populações consideradas.

Frequências observadas e esperadas

As somas

$$n_{*j} = \sum_{i=1}^N n_{ij}$$

calculam as ocorrências dos valores y_j na **amostra total**, de tamanho

$$n = \sum_i n_i$$

assim que as frequências observadas dos diferentes valores de Y na amostra total são

$$f_j = \frac{n_{*j}}{n}$$

Na **hipótese de homogeneidade**, os valores esperados das ocorrências dos valores y_j dentro da i -ésima amostra são portanto

$$n_i f_j$$

Estes valores devem ser comparados com as ocorrências observadas

$$n_{ij}$$

Flutuações e qui-quadrado esperado

O **teorema limite central** sugere que as **flutuações** das ocorrências observadas n_{ij} em torno do valor esperado $n_i f_j$, ou seja,

$$(n_{ij} - n_i f_j)^2$$

sejam da ordem das variâncias esperadas

$$n f_j(1 - f_j) \simeq n_i f_j$$

Consequentemente, esperamos que a soma dos quocientes

$$Q^2 = \sum_{ij} \frac{(n_{ij} - n_i f_j)^2}{n_i f_j}$$

seja da ordem do produto

$$Q^2 \sim MN$$

o número de classes da variável vezes o número das populações, e não muito maior.

Mais uma vez, um **teorema de Pearson** diz que, quando n é grande, a lei da variável Q^2 é bem aproximada pela lei **qui-quadrado** $\chi^2((M-1)(N-1))$, com $df = (M-1)(N-1)$ graus de liberdade.

Hipótese e alternativa.

Estatística do teste.

Nível de significância e potência do teste.

Valor de prova.

Testes T de Student .

Testes qui-quadrado de Pearson.

Testes de aderência.

Testes de independência.

Testes de homogeneidade.