Testes não-paramétricos: amostras simples

- Deste ponto até o fim da capítulo 5 deixaremos os métodos bayesianos e nos concentraremos em métodos clássicos.

- 'Testes não paramétricos' implica que 'não se supõe nenhuma distribuição', mas por vezes é necessário algum grau de suposição para fazermos progresso.

- Sobre os testes não-paramétricos temos que:
  - Estes fazem menos suposições sobre os dados. Se de fato nada se sabe sobre a distribuição subjacente *não há alternativa*.
  - Se o tamanho da amostra é pequena, provavelmente é necessário usar um teste não-paramétrico.
  - Testes não-paramétricos conseguem lidar com dados em formato não-numérico, por exemplo, posições em listas e classificações. Não há equivalente paramétrico.
  - Testes não-paramétricos podem tratar amostras de observações de populações diversas.

- Quais são os contra-argumentos? A maioria diz respeito à divisão dos dados em intervalos (*binning*). Isso é ruim pois perde-se informações portanto eficiência. De todo modo a força de testes não-paramétricos é tipicamente 10% ou menos menor que suas contrapartes paramétricas.

Testes não-paramétricos: teste do qui-quadrado

- Considere que dados observacionais/experimentais podem ser divididos em intervalos e um modelo/hipótese que prediz a população de cada bin. A estatística chi-quadrado descreve a qualidade do ajuste dos dados ao modelo.

- Se os números observados em cada intervalo k são O_i e os valores esperados pelo modelo são E_i, então essas estatística é:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

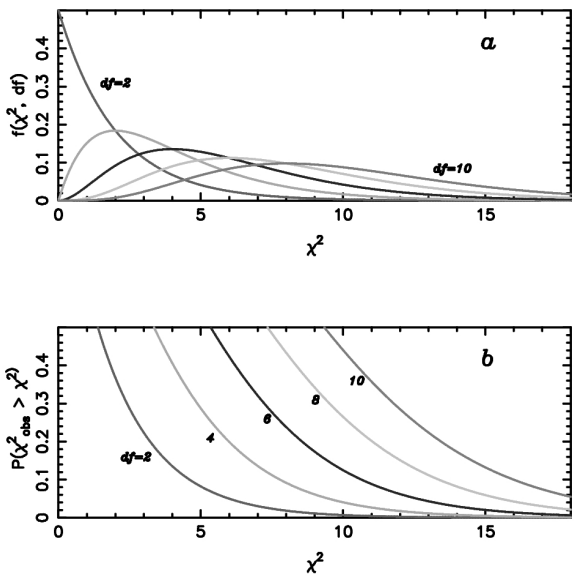
- A hipótese nula, H_0 de que o número de objetos caindo em cada categoria ou intervalo é E_i. O qui-quadrado testa se O_i está suficientemente perto de E_i para se encaixar em H_0.

Testes não-paramétricos: teste do qui-quadrado

A distribuição do qui-quadrado com $\nu = k - 1$ graus de liberdade é dada por

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

para $x \geq 0$. Que é a distribuição da variável aleatória $Y^2 = Z_1^2 + Z_2^2 + ...Z_\nu^2$, onde os Z_i são variáveis aleatórias independentes geradas por uma distribuição Normal. A Tabela B.6 apresenta os valores críticos. Se x^2 exceder esses valores então H_0 é rejeitada com o dado nível de significância.



a) A densidade de probabilidade da função χ^2 para df graus de liberdade; *b*) a função de distribuição de $\int_{\chi^2}^{\infty} f(\chi^2, df)\, d\chi^2$, que é usada para determinar se o χ^2 é grande o suficiente para rejeitar-se H_0.

Testes não-paramétricos: teste do qui-quadrado

- A premissa básica do teste qui-quadrado é de que os desvios de E_i são devidas a flutuações estatísticas do limitado número de observações por intervalo, ou seja "ruído" ou variação estatística de Poisson e o teste simplesmente dá a probabilidade dos desvios de O em relação a E ocorrerem por "azar".

- Como veremos mais adiante é necessário garantir que o número de eventos por intervalo para garantir que em cada termo da gaussiana a some seja aproximadamente gaussiana.

- O teste do qui-quadrado faz a seguinte pergunta: "Qual é a probabilidade deste valor que χ^2 ou um maior ainda ocorrerem apenas por azar?". Por isso é usada a distribuição cumulativa da figura anterior, lembrando que a probabilidade de qualquer valor exato do χ^2 é infinitamente pequena.

Testes não-paramétricos: teste do qui-quadrado

- Boas notícias sobre o tese do qui-quadrado:

  1. É conhecido e bem aceito na comunidade científica
  2. Como o χ^2 é aditivo, ele pode ser usado para testar amostras diferentes, com intervalos diferentes, que são previstos pelo mesmo modelo, todas ao mesmo tempo.
  3. O χ^2 é fácil de ser computado e de ser interpretado, pois a média do χ^2 é igual ao número de graus de liberdade e sua variância é igual a 2ν. Ou seja, aproximadamente, para um número de bins ≥ 4, se o valor de χ^2 for da ordem do número de intervalos-1, então aceite H_0, se for o dobro disso, você provavelmente rejeitará H_0. Por essa razão é comum usarmos o *qui-quadrado reduzido* (χ^2/ν). Se esse for da ordem da unidade aceita-se H_0, se muito maior, rejeita-se.
  4. A minimização do χ^2 é um método muito comum de ajuste de modelos, como veremos no próximo capítulo.

Testes não-paramétricos: teste do qui-quadrado

- Más notícias sobre o tese do qui-quadrado:

  1. Os dados tem que ser divididos em intervalos (*bins*), o que tira um pouco da força do teste e resolução dos dados.

  2. Cada intervalo deve ter um certo tamanho que evite as instabilidades derivadas de $E_i \rightarrow 0$. Isso nos leva a outra regra informal: mais de 80% dos intervalos devem ter $E_i > 5$. Se necessário, intervalos devem ser combinado para garantir que isso se verifique.

  3. O teste de qui-quadrado não dá informação a respeito de *direção*, ou seja é um teste 'bi-lateral'. Ele diz apenas se as diferenças entre dados e modelos podem ser devidas a flutuações estatísticas apenas.

- Em vista desses problemas é de se suspeitar que deve existir algo melhor que isso, e da fato há.

Testes não-paramétricos: teste de uma amostra de Kolmogorov-Smirnov

- Esse é um teste muito simples de ser executado:

  1. Calcule $S_e(x)$, a frequência cumulativa *predita* pelo modelo sob H_0.
  2. Considere uma amostra de N observações e calcule $S_o(x)$, a distribuição cumulativa *observada*, ou seja a soma de todas as observações de cada x divididos pela some de todas as N observações.
  3. Encontre
  $$D = max|S_e(x) - S_o(x)|$$
  4. Consulte uma tabela como B.7 com a distribuição conhecida de D sob H_0 e determine o destino da hipótese nula. Se D exceder o valor crítico para o N apropriado, então H_0 pode ser rejeitado dentro do nível de significância dado.

- Bem como o teste de qui-quadrado a distribuição amostra indica se a divergência entre dados e modelo é 'razoável' ou seja, possível de existir somente por contsa de flutuações estatísticas.

Testes não-paramétricos: teste de uma amostra de Kolmogorov-Smirnov

$$D = max|S_e(x) - S_o(x)|$$

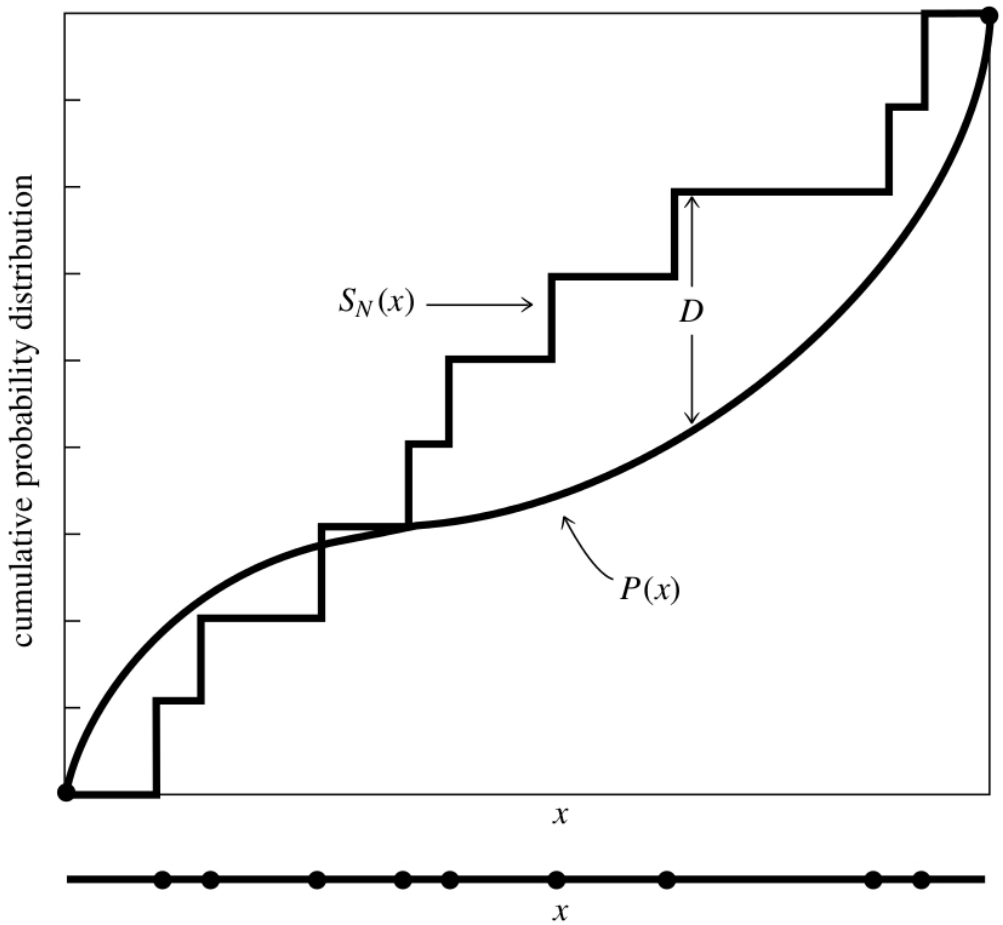


Figure 14.3.1. Kolmogorov-Smirnov statistic D. A measured distribution of values in x (shown as N dots on the lower abscissa) is to be compared with a theoretical distribution whose cumulative probability distribution is plotted as $P(x)$. A step-function cumulative probability distribution $S_N(x)$ is constructed, one that rises an equal amount at each measured point. D is the greatest distance between the two cumulative distributions.

Testes não-paramétricos: teste de uma amostra de Kolmogorov-Smirnov

- O teste K-S tem várias vantagens em relação ao χ^2. Primeiramente ele não perde informação por conta de agrupamento. Segundo ele funciona bem para amostras pequenas. De fato para amostras muito pequenas ele é a única alternativa e para amostras médias o mais poderoso. Finalmente, o teste K-S, tal qual ele foi descrito aqui é também bi-lateral, mas é possível ajusta-lo para que possibilite a determinação da *direção*.

- Porque então ele não é tão usado? Primeiramente porque as distribuições tem que ser contínuas da variável (x) para o teste ser aplicável, enquanto que o teste de χ^2 é aplicável a dados não numéricos, desde que possam ser agrupados en categorias.

- A segunda razão é que o teste de χ^2 é facilmente adaptável para o ajuste de modelos ao simplesmente reduzir-se o número de graus de liberdade de acordo com o número de parâmetros do modelo. O teste K-S não pode ser adaptado do mesmo modo, já que a distribuição de D não é conhecida quando parâmetros da população são estimados a partir da amostra.

Testes não-paramétricos: teste de aleatoriedade de uma amostra ou dos *runs*

- Esse simples teste reside em formar uma estatística binária (0-1) a partir dos dados, por exemplo cara ou coroa ou o sinal dos resíduos sobre uma média ou modelo ajustado.

- Ele serve para testar se a amostra é aleatória ou não, ou seja, se sucessivas observações são independentes.

- O teste dos runs baseia-se na análise de uma sequência de dois tipos de símbolos, digamos, A e B. Um *run* e uma subsequência de símbolos iguais. Por exemplo, a sequência

  $$ABAABBBBAAAAABBABAAABABBB$$

  tem 12 *runs*

  $$A\ B\ AA\ BBBB\ AAAAA\ BB\ A\ B\ AAA\ B\ A\ BBB$$

  valores muito pequenos ou muito grandes do número de *runs* r, levam a rejeição da hipótese nula (aleatoriedade) [1]

- O caso mais geral é que o número de *runs* seja menor do que o esperado (situação que gerem um número demasiadamente alto de *runs* são infrequentes). Desse modo temos um teste lateral (*one-tailed*).

[1] Trecho retirado de `http://www.mat.uc.pt/~cmtm/ECwww/TestesNP.pdf` de Cristina Maria Tavares Martins

Testes não-paramétricos: teste de aleatoriedade de uma amostra ou dos *runs*

- Dados, por exemplo, N jogadas de moeda que resultam em n 'caras' e m 'coroas', temos que o valor esperado do número de *runs* é

$$\mu_r = \frac{2mn}{N} + 1$$

onde, obviamente, $N = n + m$. Para grandes N's essa aproximação é assintoticamente gaussiana com

$$\sigma_r = \sqrt{\frac{2mn(2\ nm - N)}{N^2(N-1)}}$$

- Para grandes amostras, então, é possível o uso da distribuição Normal na forma padrão usando

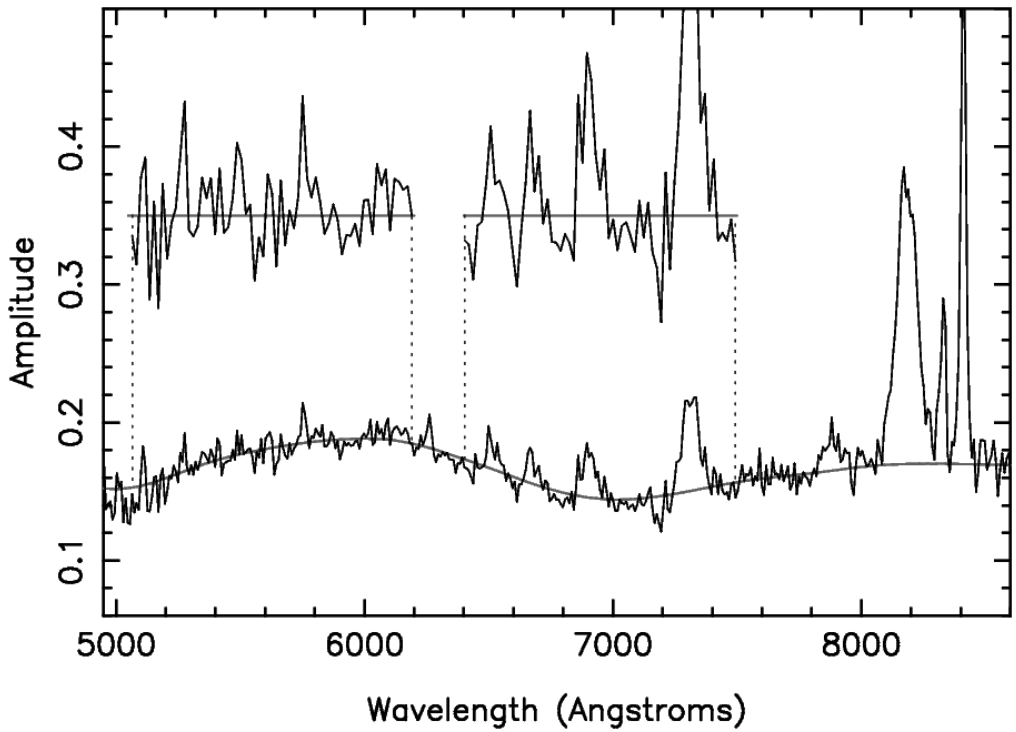
$$z = \frac{r - \mu_r}{\sigma_r}$$

e consultando a Tabela B1 a função gaussiana integral ou a função erro (*erf*). Esse é o procedimento quando $N > 20$ e chega-se ao limite da tabela B.8.

Testes não-paramétricos: teste de aleatoriedade de uma amostra ou dos *runs*

-

Exemplo W&J P. 112



Espectro do quasar 3C207. A linha cinza é um ajuste do contínuo feita através do método de mínimas componentes de Fourier. As regiões consideradas para o teste de *runs* podem ser vistas nas ampliações, cada qual com seu contínuo subtraído.

- Esse teste é muito potente para testar a independência de dados subsequentes, por exemplo, num espectro ou sequência temporal.

Testes não-paramétricos de duas amostras: teste exato de Fischer

- Nessa sub-seção lidaremos com duas amostras e queremos saber se elas foram geradas a partir da mesma distribuição ou não e, em caso negativo, se eles diferem numa direção previsível. Como não sabemos nada sobre suas distribuições temos que usar métodos não-paramétricos.

- Esse teste é para duas amostras independentes e *pequenas*, para as quais se tem dados *binários* discretos, por exemplo, a pontuação da tabela abaixo, onde as categorias são mutuamente excludentes:

Amostra =	1	2
Categoria = 1	A	C
= 2	B	D

- Calcule a seguinte estatística:

$$p = \frac{(A+B)!\,(C+D)!\,(A+C)!\,(B+D)!}{N!A!B!C!D!}$$

essa é a probabilidade de do total de $N = A + B + C + D$ pontos podem ser da forma como são *quando as duas amostras são de fato idênticas*.

- De fato o teste faz a seguinte pergunta: qual é a probabilidade da ocorrência de um dados resultado ou um mais extremos sob H_0 (hipótese das pontuações serem aleatórias).

Testes não-paramétricos de duas amostras: teste de qui-quadrado para duas amostras

- Aqui também o teste de chi-quadrado é aplicável. Com todas as preocupações já mencionadas pode ser a única alternativa para dados não numéricos.
- Para começar cada amostra é dividida nas mesmas r divisões (*bins*; uma tabela de $k \times r$ valores). A hipótese H_0 é que todos as k amostras tem a mesma população.

Amostra: $j =$		1	2	3
Categoria: $i =$	1	O_{11}	O_{12}	O_{13}
	2	O_{21}	O_{22}	O_{23}
	3	O_{31}	O_{32}	O_{33}
	.	...	...	...

- Calcule então:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- E_{ij} são os valores esperados calculados a partir de

$$E_{ij} = \frac{\sum_{i=1}^{r} O_{ij} \sum_{j=1}^{k} O_{ij}}{\sum_{i=1}^{r} \sum_{j=1}^{k} O_{ij}}$$

Sob H_0, isso é distribuído como χ^2 com $(r-1)(k-1)$ graus de liberdade.

Testes não-paramétricos de duas amostras: teste de qui-quadrado para duas amostras

- Note que há uma modificação nesse teste para o caso da tabela de pontuação 2×2, com um total de N objeto. Nesse caso:

$$\chi^2 = \frac{N(|AD - BC| - N/2)^2}{(A+B)(C+D)(A+C)(B+D)},$$

que tem apenas um grau de liberdade.

- Os cuidados usuais com o χ^2 são os mesmos já citados - cuidado com a número de contagens menor que 5 em cada célula. Caso isso ocorra, funda células, simule a distribuição da estatística teste sob a hipótese nula ou abandone o teste. No caso de 2×2 células o número total deve ser maior do que 30, senão use o teste exato de Fisher.

Testes não-paramétricos: teste U

- Esse teste é usualmente preferível ao qui-quadrado, principalmente porque evita a divisão em células ou intervalos.
- Há duas amostras, A (com m membros) e B (n membros). H_0 é que A e B são gerados a partir da mesma distribuição, enquanto que H_1 pode ser:
  - que A é estocasticamente maior do que B (i.e. $Prob(A > x) > Prob(B > x)$)
  - que B é estocasticamente maior do que A
  - que A e B sejam diferentes em algum outro modo, talvez em espalhamento ou *skewness*.
- As primeira duas hipótese são direcionais, resultando em testes uni-laterais. A terceiro é o caso oposto.

Testes não-paramétricos: teste U

- Para prosseguir escolha H_1 e uma valor de α (probabilidade de rejeitar H_0 sendo essa verdadeira) e faça:
  - Ordene em ordem ascendente a amostra combinada A+B, preservando a informação de quem pertence a qual amostra.
  - (Dependendo da escolha de H_1) some o número de posições de elemento de A para obter U_A. Faça o mesmo para B para obter U_B. No caso de empates atribua o valor da médias das posições empatadas. de modo que apenas uma soma é necessária para determinar a ambos.
  - A distribuição dos valores de U é conhecida e está na tabela B.9. Nas colunas marcadas como c_u (*upper-tail probabilities*) estão os valores das probabilidades de ocorrência de valores de U como o observado ou maior, sob H_0

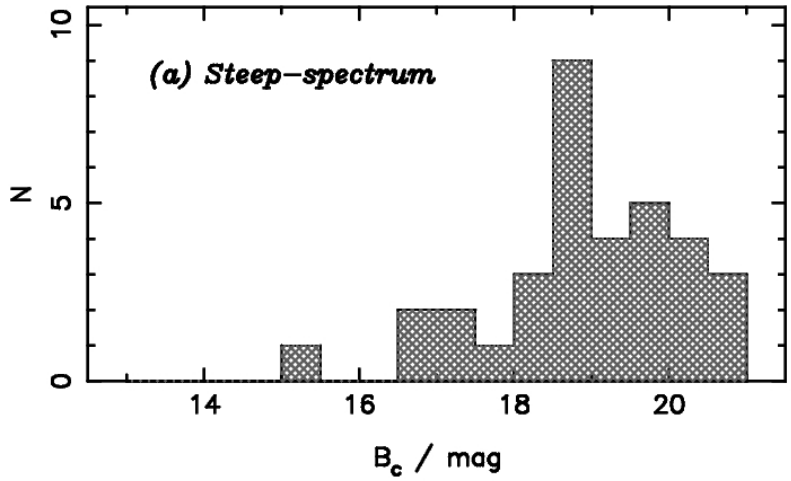
- A tabela apresenta valores apenas para $m \leq 4$ e $n \leq 10$. Para amostras até $m = 10$ e $n = 12$ veja Siegel & Castellan (1988). Para amostras ainda maiores a distribuição de U_A tende a uma Normal com $\mu_A = m(N+1)/2$ e $\sigma_A^2 = mn(N+1)/12$, onde a significância pode ser calculada por

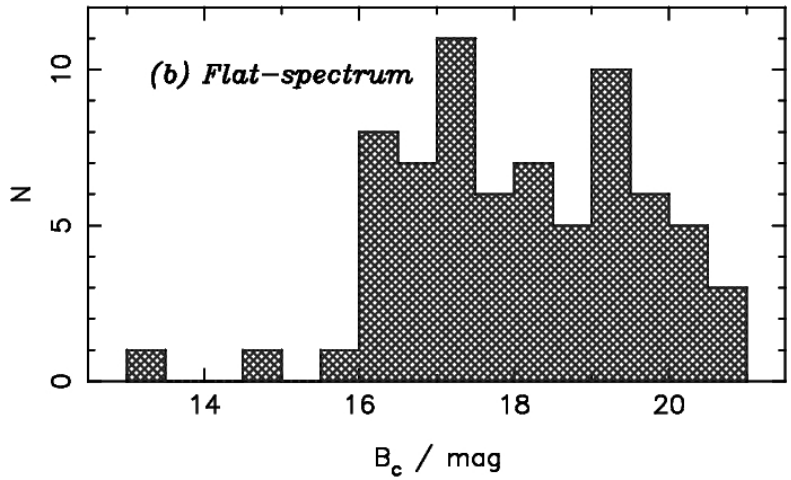
$$z = \frac{U_A \pm 0,5 - \mu_A}{\sigma_A},$$

onde o sinal de $\pm 0,5$ é positivo para as probabilidades de $U \leq$ o observado e negativo no caso oposto. No caso bi-lateral (as amostras são distinguíveis) simplesmente dobre as probabilidades determinadas nas tabelas.

Testes não-paramétricos: teste U

Exemplo W&J P. 116





Testes não-paramétricos de duas amostras: teste K-S para duas amostras

- A formulação é a mesma do teste K-S para uma amostra; considera o máximo desvio entre as distribuições cumulativas normalizadas das duas amostras com m e n membros. Como sempre H_0 é a hipótese de ambas pertencerem à mesma população. H_1 pode ser a hipótese delas diferirem (teste bi-lateral) ou de diferirem numa dada direção (teste uni-lateral).

- Para implementar o testes veja o procedimento para o teste de uma amostra e apenas substitua S_e e S_o por S_m e S_n, correspondente às duas amostras.

- Valores críticos de D são dados nas tabelas B.10 (*one-tailed*) e B.11(*two-tailed*). Para amostras grandes (*one-tailed*) calcule:

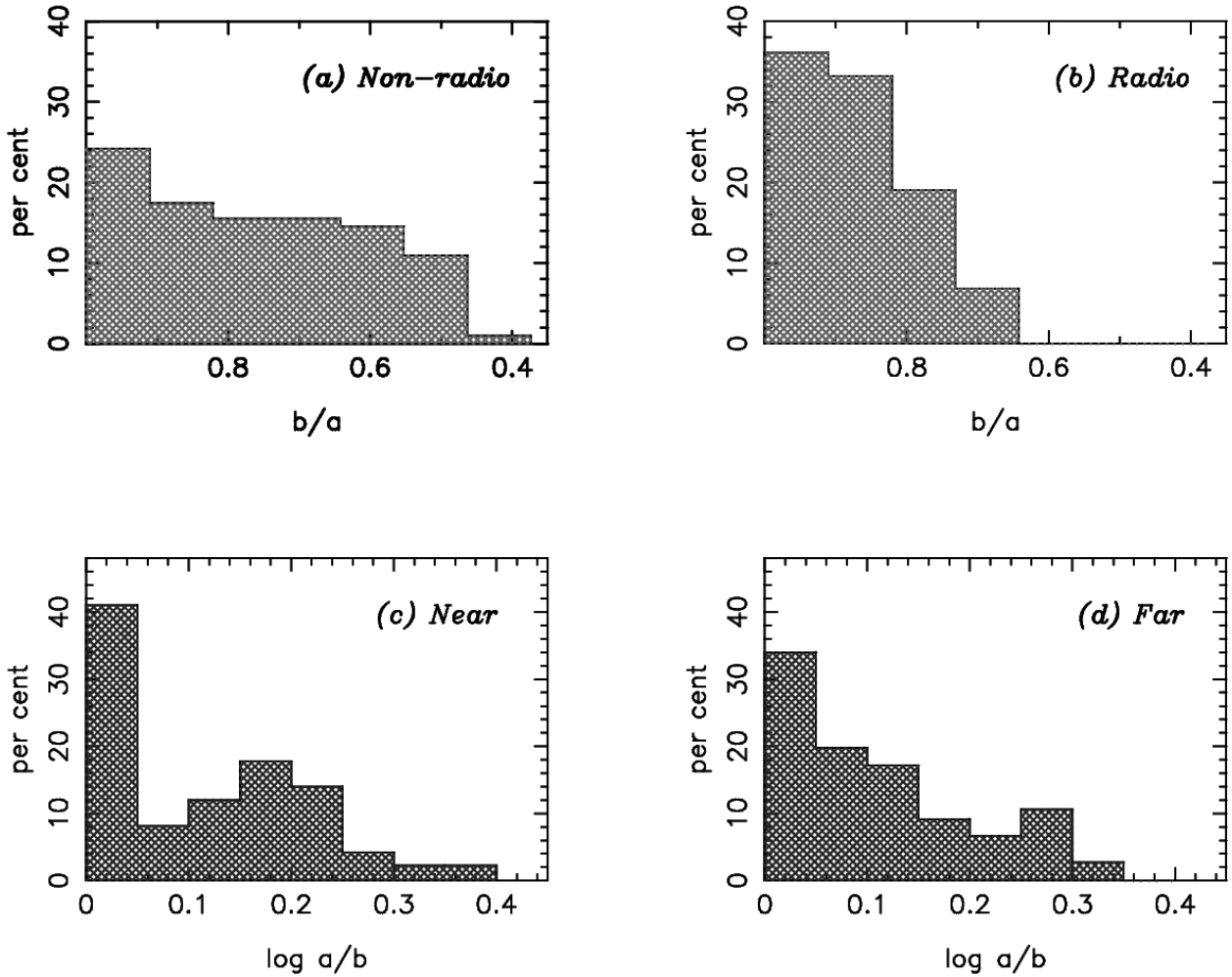
$$\chi^2 = 4D^2 \frac{mn}{m+n}$$

que segue uma distribuição aproximada de chi-quadrado com dois graus de liberdade.

- Essa teste é muito mais eficiente que o teste t, principalmente para amostras pequenas. Suas eficiência sempre supera a do teste de qui-quadrado e supera levemente o teste U para amostras pequenas. Para grandes amostras o teste U é o mais eficiente.

Testes não-paramétricos: teste U

Exemplo W&J P. 118



Razão entre o semi-eixo menor e maior de galáxias elípticas.

Testes não-paramétricos de duas amostras: teste K-S para duas amostras

- As tabelas a seguir (adaptadas de Siegel & Castellan 1988) tentam resumir o, aparentemente, vasto mundo dos testes não paramétricos. É realmente tão vasto assim? Para decidir que testes devem ser usado os pontos a seguir devem ser considerados na sua decisão.

  - Os casos de duas ou K amostras contém testes para *amostras relacionadas*, ou seja, amostras emparelhadas. Isso é uma prática experimental comum em biologia e ciências do comportamento onde o conceito de *amostras de controle* é muito bem desenvolvido. Esse não é o caso da astronomia por razões óbvias, mas esses é usado eventualmente. A Tabela 40 mostras os testes mais poderosos para tratar tais experimentos.

  - Na Tabela 40 os testes vão sendo mais sofisticados em termos de medidas, a medida que descemos. Vamos de *Nominal* (na qual os objetos são simplesmente colocados em conjuntos ou intervalos) a *Ordinal* (no qual os objetos são ordenados ou ranqueados) até *Intervalar* (na qual os objetos são colocados numa escala , não necessariamente numérica, na qual a distância na escala é relevante). Nenhum desses testes requerem medidas numa escala de *Razão*, o caso mais forte de escala de medida na qual além das propriedades da intervalar, um ponto-zero verdadeiro é adicionado. Por fim a tabela é cumulativa de cima para baixo. Para um dado nível de medida todos os testes acima são aplicáveis.

  - A eficiência de um teste depende muito da sua aplicação. As tabelas 41 e 42 podem ajudar no processo de escolha.

Table 5.4 *Non-parametric tests for comparison of samples*

Level of measurement	One-sample case	Two-sample case		k-sample case	
		Related	Independent	Related	Independent
Nominal or categorical	Binomial test *chi-square test	McNemar change test	*Fisher exact test for 2×2 tables *chi-square test for $r \times 2$ tables	Cochran Q test	*chi-square test for $r \times k$ tables
Ordinal or ordered	*Kolmogorov–Smirnov one-sample test *One-sample runs test Change-point test	Sign test Wilcoxon signed-ranks test*	Median test *U (Wilcoxon–Mann–Whitney) test Robust rank-order test *Kolmogorov–Smirnov two-sample test Siegel–Tukey test for scale-differences	Friedman two-way analysis of variance by ranks Page test for ordered alternatives	Extension of median test Kruskal–Wallis one-way analysis of variance Jonckheere test for ordered alternatives
Interval		Permutation test for paired replicates	Permutation test for two independent samples Moses rank-like test for scale differences		

*Described in this chapter; Siegel & Castellan (1988) discuss the other tests.

Table 5.5 *Single-sample non-parametric tests*

Test	Applicability[2]	$N < 10$?	Comment
Binomial test	Goodness-of-fit (N)	Yes	Appropriate for two-category (dichotomous) data; do *not* dichotomize continuous data.
[1]Chi-square test	Goodness-of-fit (N)	No	For testing categorized, pre-binned, or classified data; choose categories with expected frequencies 6–10.
[1]Kolmogorov–Smirnov one-sample test	Goodness-of-fit (O)	Yes	The most powerful test for data from a continuous distribution; may always be more efficient than chi-square test.
[1]One-sample runs test	Randomness of event sequences (O)	Yes	Does not estimate differences between groups.
Change-point test	Change in the distribution of an event sequence (O)	Yes	Robust with regard to changes in distributional form; efficient.

[1] Described in this chapter; Siegel & Castellan (1988) discuss the other tests.
[2] *Goodness-of-fit* indicates general testing for any type of difference, i.e. H_o is that the distribution is drawn from the specified population. The level of measurement required is indicated by N – Nominal, O – Ordinal, or I – Interval.

Table 5.6 *Two-sample non-parametric tests*

Test	Applicability[2]	$N < 10$?	Comment
[1]Fisher exact test for 2×2 tables	Difference (N)	Yes	The most powerful test for dichotomous data.
[1]Chi-square test for $r \times 2$ tables	Difference (N)	No	Best for pre-binned, classified, or categorized data.
Median test	Location (O)	Yes	Best for small numbers; efficiency *decreases* with N.
[1]U (Wilcoxon–Mann–Whitney) test	Location (O)	Yes	One of the most efficient non-parametric tests.
Robust rank-order test	Location (O)	Yes	Efficiency similar to U test.
[1]Kolmogorov–Smirnov two-sample test	Two-tailed: Difference One-tailed: Location (O)	Yes	The most powerful test for data from a continuous distribution.
Siegel–Tukey test for scale-differences	Dispersion (O)	Yes	The medians must be the same (or known) for both distributions. Low efficiency.
Permutation test	Location (I)	Yes	Very high efficiency.
Moses rank-like test for scale-differences	Dispersion (I)	(No)	Does not require identical medians; valid for small samples, but efficiency increases with sample size.

[1] Described in this chapter; Siegel & Castellan (1988) discuss the other tests.
[2] *Difference* signifies sensitivity to any form of difference between the two distributions, i.e. H_o is that the two distributions are drawn from the same population; *Location* indicates sensitivity to the position of the distributions, e.g. means or medians; and *Dispersion* indicates sensitivity to the spread of the distributions, i.e. variance, rms extremes. The level of measurement required is indicated by N – Nominal, O – Ordinal or I – Interval.