# X. Testování hypotéz: neparametrické testy

# Co přináší neparametrické testování hypotéz

V případě ordinálních (pořadových) nebo nominálních dat odpovídá na specifické otázky:

1. Existuje významný soulad dané charakteristiky rozdělení četnosti vzorku se zadanou charakteristikou populace?

2. Existuje významný rozdíl dané charakteristiky mezi 2 (nebo více) vzorky?

Charakteristika - např. medián, zadané pořadí, **typ rozdělení pr-sti (četnosti)** aj.

# Neparametrické testy hypotéz

- Má medián populace s neznámým rozdělením stanovenou hodnotu? (**mediánový test**)
- Pochází výběr z populace se zadaným (známým) rozdělením pravděpodobnosti? (**Chi-kvadrát test**)

# Mediánový test

(pro 1 výběr)

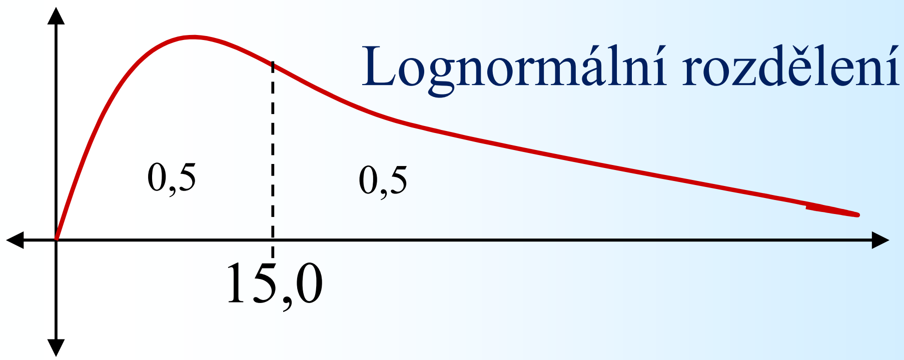
- Nevíme-li, zda má populace normální rozdělení, předpokládáme, že má medián $\tilde{\mu}_0$ rozsah vzorku $n$
- $H_0$: $\tilde{\mu} = \tilde{\mu}_0$, $H_1$: $\tilde{\mu} \neq \tilde{\mu}_0$ - oboustranný test
- Testové kritérium: $u = \dfrac{|2m - n|}{\sqrt{n}}$

$m$ je počet počet pozorování ve vzorku < $\tilde{\mu}_0$

- Jestliže $u > z_{1-\alpha/2}$ potom $H_0$ zamítáme!

$z_{1-\alpha/2}$ je kvantil norm. normál. rozd. (viz tabulky)

# **Příklad 1**: Mzdy




Náhodně vybraný vzorek 19 pracovníků jisté (dělnické) profese ve městě Karviná poskytl následující údaje o jejich měsíčních mzdách (v tis.Kč):

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10,0 | 12,3 | 12,6 | 12,6 | 13,0 | 13,2 | 13,3 | 13,3 | 13,4 | 13,8 |
| 14,1 | 14,3 | 14,6 | 15,1 | 15,2 | 15,4 | 16,5 | 18,2 | 20,5 | ---- |

Na hladině významnosti $\alpha = 0{,}05$ testujte hypotézu, že průměrná (mediánová) měsíční mzda pracovníků této profese v Karviné je 15 tis. Kč.

## **Příklad 1**: Řešení …

**Populace** - měsíční mzdy všech pracovníků dané profese v Karviné

Je známo, že **mzdy nemají normální rozdělení** pr-sti!

Proto namísto **střední hodnoty** je lepší

charakteristikou **medián**, jemu pak odpovídá

neparametrický dvoustranný mediánový test hypotézy

$$H_0: Med(X) = 15$$

proti alternativní hypotéze

$$H_1: Med(X) \neq 15$$

## **Příklad 1**: … Řešení

Z dat: $n = 19$, $m = 13$, vypočteme:

$$u = \frac{|2.13 - 19|}{\sqrt{19}} = 1{,}61$$

NORMSINV(0,975) = 1,96

Protože 1,61 < 1,96, nulovou hypotézu $H_0$ nezamítáme (přijímáme)

**Jinými slovy:** na zvolené hladině významnosti 0,05 vzorek neodporuje hypotéze o výši mediánové měsíční mzdy prac. dané profese v Karviné (tj. 15 tis. Kč)

**Také**: vybraný vzorek je v souladu s karvinskou populací v této profesi!

# Chi-kvadrát test

($X^2$ - test pro 1 výběr)

- Data mohou být **nominální** (nejslabší požadavek)!
- Testuje se (nulová) hypotéza
  $H_0$: výběr pochází z populace s daným rozdělením
- Zadané rozdělení je obvykle:

- diskrétní rozdělení se stejnými pr- stmi (tzv. **test nezávislosti**)

- diskrétní rozdělení s rozdílnými pr- stmi (tzv. **test dobré shody**)

## **Příklad 2**: Limonády

Nová limonáda se prodávala za stejnou cenu jeden týden ve 3 různých typech obalu: A, B, C, počet prodaných limonád viz tabulka:

| Typ obalu | Prodané kusy |
|---|---:|
| A | 135 |
| B | 130 |
| C | 155 |
| Celkem | 420 |

Ovlivňuje styl designu obalu počet prodaných limonád?

**Jinak**: Závisí prodej na obalu?

## **Příklad 2**: Algoritmus a řešení 1: ***Test nezávislosti***

**Krok 1**. **Nulová hypotéza $H_0$:**

Počet prodaných kusů **nezávisí** na typu obalu (rozdíly v prodeji u vzorku jsou pouze dílem náhody).

**Očekávané četnosti (Expected):**

$E_1 = E_2 = E_3 = {}^{420}/_{3} = 140$

**Pozorované četnosti (Observed):**

$O_1 = 135,\ O_2 = 130,\ O_3 = 155$

**Krok 2. Testové kritérium:** $$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$k$ **- počet kategorií** ($k = 3$)

# **Příklad 2**: Algoritmus a řešení 2

**Krok 3**. Porovnání hodnoty vypočítaného kritéria

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 2{,}5$$

$\boxed{\text{CHIINV}(0{,}05;2) = 6{,}0}$

s tabulkovou **kritickou hodnotou** rozdělení $\chi^2_\alpha(2) = 6{,}0$

kde $\alpha\,( = 0{,}05)$ je zadaná hladina významnosti

**V každé kategorii: $O_i$ alespoň 5 !**

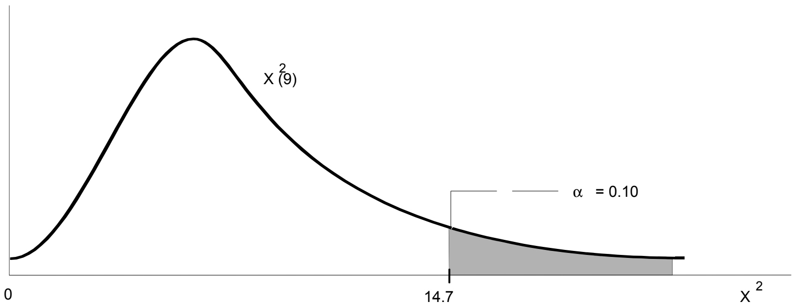
Jestliže

$$X^2 = 2{,}5 < \chi^2_{0{,}05}(2) = 6{,}0$$

potom $H_0$ nezamítáme! (jinak **zamítáme**)

***p*-hodnota (signifikance)** = 0,287 > 0,05 (Nezamítáme)

Kritické hodnoty
rozdělení Chi-kvadrát $\chi^2_\alpha(n)$




| n \ α | 0,995 | 0,99 | 0,975 | 0,95 | 0,9 | 0.1 | 0.05 | 0.025 | 0.01 | 0,005 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 0,00 | 0,00 | 0,00 | 0,02 | 2.7 | 3.8 | 5.0 | 6.6 | 7.9 |
| 2 | 0,01 | 0,02 | 0,05 | 0,10 | 0,21 | 4.6 | 6.0 | 7.4 | 9.2 | 10.6 |
| 3 | 0,07 | 0,12 | 0,22 | 0,35 | 0,58 | 6.3 | 7.8 | 9.4 | 11.3 | 12.8 |
| 4 | 0,21 | 0,30 | 0,48 | 0,71 | 1,06 | 7.8 | 9.5 | 11.1 | 13.3 | 14.9 |
| 5 | 0,41 | 0,55 | 0,83 | 1,15 | 1,61 | 9.2 | 11.1 | 12.8 | 15.1 | 16.7 |
| 6 | 0,68 | 0,87 | 1,24 | 1,64 | 2,20 | 10.6 | 12.6 | 14.4 | 16.8 | 18.5 |
| 7 | 0,99 | 1,24 | 1,69 | 2,17 | 2,83 | 12.0 | 14.1 | 16.0 | 18.5 | 20.3 |
| 8 | 1,34 | 1,65 | 2,18 | 2,73 | 3,49 | 13.4 | 15.5 | 17.5 | 20.1 | 22.0 |
| 9 | 1,74 | 2,09 | 2,70 | 3,33 | 4,17 | 14.7 | 16.9 | 19.0 | 21.7 | 23.6 |
| 10 | 2,16 | 2,56 | 3,25 | 3,94 | 4,87 | 16.0 | 18.3 | 20.5 | 23.2 | 25.2 |
| 11 | 2,60 | 3,05 | 3,82 | 4,57 | 5,58 | 17.3 | 19.7 | 21.9 | 24.7 | 26.8 |
| 12 | 3,07 | 3,57 | 4,40 | 5,23 | 6,30 | 18.5 | 21.0 | 23.3 | 26.2 | 28.3 |
| 13 | 3,57 | 4,11 | 5,01 | 5,89 | 7,04 | 19.8 | 22.4 | 24.7 | 27.7 | 29.8 |
| 14 | 4,07 | 4,66 | 5,63 | 6,57 | 7,79 | 21.0 | 23.7 | 26.1 | 29.1 | 31.3 |
| 15 | 4,60 | 5,23 | 6,26 | 7,26 | 8,55 | 22.3 | 25.0 | 27.5 | 30.6 | 32.8 |
| 16 | 5,14 | 5,81 | 6,91 | 7,96 | 9,31 | 23.5 | 26.3 | 28.8 | 32.0 | 34.3 |
| 17 | 5,70 | 6,41 | 7,56 | 8,67 | 10,09 | 24.8 | 27.6 | 30.2 | 33.4 | 35.7 |
| 18 | 6,26 | 7,01 | 8,23 | 9,39 | 10,86 | 26.0 | 28.9 | 31.5 | 34.8 | 37.2 |
| 19 | 6,84 | 7,63 | 8,91 | 10,12 | 11,65 | 27.2 | 30.1 | 32.9 | 36.2 | 38.6 |
| 20 | 7,43 | 8,26 | 9,59 | 10,85 | 12,44 | 28.4 | 31.4 | 34.2 | 37.6 | 40.0 |
| 21 | 8,03 | 8,90 | 10,28 | 11,59 | 13,24 | 29.6 | 32.7 | 35.5 | 38.9 | 41.4 |
| 22 | 8,64 | 9,51 | 10,98 | 12,34 | 14,04 | 30.8 | 33.9 | 36.8 | 40.3 | 42.8 |
| 23 | 9,26 | 10,20 | 11,69 | 13,09 | 14,58 | 32.0 | 35.2 | 38.1 | 41.6 | 42.2 |
| 24 | 9,89 | 10,86 | 12,40 | 13,85 | 15,66 | 33.2 | 36.4 | 39.4 | 43.0 | 45.6 |
| 25 | 10,52 | 11,52 | 13,12 | 14,61 | 16,47 | 34.4 | 37.7 | 40.6 | 44.3 | 46.9 |
| 26 | 11,16 | 12,20 | 13,84 | 15,38 | 17,29 | 35.6 | 38.9 | 41.9 | 45.6 | 48.6 |
| 27 | 11,81 | 12,88 | 14,57 | 16,15 | 18,11 | 36.7 | 40.1 | 43.2 | 47.0 | 49.6 |
| 28 | 12,46 | 13,56 | 15,31 | 16,93 | 18,94 | 37.9 | 41.3 | 44.5 | 48.3 | 51.0 |
| 29 | 13,12 | 14,26 | 16,05 | 17,71 | 19,77 | 39.1 | 42.6 | 45.7 | 49.6 | 52.3 |
| 30 | 13,79 | 14,95 | 16,79 | 18,49 | 20,60 | 40.3 | 43.8 | 47.0 | 50.9 | 53.7 |

# Řešení příkladu 2 pomocí Excelu:

Tabulka→Funkce: SUMA, CHIIV, CHIDIST…

| Typ obalu | Oi | Ei | (Oi - Ei)^2/Ei |
|---|---|---|---|
| A | 135 | 140 | 0,179 |
| B | 130 | 140 | 0,714 |
| C | 155 | 140 | 1,607 |
| Sumy | 420 | 420 | 2,500 |

| | |
|---|---|
| X^2 = | 2,5 |
| alfa = | 0,05 |
| k-1 = | 2 |
| CHIINV = | 5,991 |
| CHIDIST = | 0,287 |

**Signifikance** = CHIDIST = 0,287 > 0,05

$\Rightarrow H_0$ nezamítáme!

# **Příklad 2**: Limonády (**nová verze**)…

Nová limonáda se prodávala za stejnou cenu jeden týden ve fakultním bufetu ve 3 různých typech obalu: A, B, C, počet prodaných limonád viz tabulka:

NOVÉ ZADÁNÍ:

| Typ obalu | Prodané kusy | NOVÉ ZADÁNÍ |
|---|---|---|
| A | 135 | 120 |
| B | 130 | 130 |
| C | 155 | 170 |
| Celkem | 420 | 420 |

Ovlivňuje styl designu obalu počet prodaných limonád?

# Řešení příkladu 2… pomocí Excelu:

Tabulka→Funkce: SUMA, CHIIV, CHIDIST…

| | Oi | Ei | (Oi-Ei)^2/Ei |
|---|---|---|---|
| A | 120 | 140 | 2,857 |
| B | 130 | 140 | 0,714 |
| C | 170 | 140 | 6,429 |
| Suma | 420 | 420 | 10,000 |

$$\text{Signifikance} = \text{CHIDIST} = 0{,}0067 < 0{,}05 \Rightarrow \text{H}_0$$

**zamítáme**!

$$X^2 = 10{,}0 > \chi^2_\alpha(k-1) = \text{CHINV} = 5{,}991 \Rightarrow \text{H}_0$$

zamítáme!

# **Příklad 3**: Barvy automobilů 1

Automobil Škoda - Felicia se prodává ve čtyřech barvách:

- 40% zákazníků požaduje zelenou barvu automobilu
- 25% červenou barvu,
- 25% modrou barvu a
- 10% bílou barvu.

K ověření správnosti předpokladu o struktuře poptávky podle barev použijte záznamy o nákupech v dané prodejně v jistém měsíci

# **Příklad 3**: Barvy automobilů 2

Vstupní údaje obsahuje následující tabulka:

| j | Barva | $p_{0,j}$ | $n_j$ |
|---|---|---|---|
| 1 | zelená | 0,40 | 201 |
| 2 | červená | 0,25 | 105 |
| 3 | modrá | 0,25 | 144 |
| 4 | bílá | 0,10 | 30 |
| součet | | 1,00 | 480 |

Na hladině významnosti $\alpha = 0{,}05$ testujte hypotézu, že uvedené pravděpodobnostní odhady odpovídají zjištěným hodnotám prodejů

# Příklad 3: Algoritmus a řešení 1: *Test dobré shody*

**Krok 1**. **Nulová hypotéza $H_0$:**

$$H_0 : p_{0,1} = 0{,}4, p_{0,2} = p_{0,3} = 0{,}25, p_{0,4} = 0{,}1$$

**Očekávané četnosti**:

$E_1 = 192$, $E_2 = 120$, $E_3 = 120$, $E_4 = 48$

**Pozorované četnosti**:

$O_1 = 201$, $O_2 = 105$, $O_3 = 144$, $O_4 = 30$

**Krok 2. Testové kritérium**:

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

$k$ - **počet kategorií** ($k = 4$)

## Očekávané četnosti:

$$\text{Očekáv\_čet\_}i = \text{Pravděp\_}i \times \text{celk\_čet}$$

**Příklad:**

$i$ = zelená, $\text{Pravděp\_}i = 0{,}40$ , $\text{celk\_čet} = 480$

$E_1 = \text{Očekáv\_čet\_}i = 0{,}4 * 480 = 192$

atd.

# Příklad 3: Algoritmus a řešení 2: *Test dobré shody*

**Krok 3**. Porovnání hodnoty vypočítaného kritéria

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 13{,}85$$

s tabulkovou **kritickou hodnotou** rozdělení $\chi^2_{0,05}(3) = 7{,}81$

**V každé kategorii: *$O_i$* je alespoň 5** ( >30)

Platí

$$X^2 = 13{,}85 > \chi^2_{0,05}(3) = 7{,}81 = \mathrm{CHIINV}(0{,}05;3)$$

proto $H_0$ zamítáme! Alternativně:

Sig = CHIDIST(13,85; 3) = 0,003 < 0,05

# Testování nezávislosti kvalitativních znaků 1

V jednom vzorku (výběru) můžeme současně sledovat dva nebo i více (kvalitativních) znaků

**Příklad**:

Při kontrole jakosti výrobku sledujeme přítomnost nebo nepřítomnost vady $A$ (znak $A$), nebo přítomnost nebo nepřítomnost vady $B$ (znak $B$).

$A$ i $B$ nabývají pouze dvě alternativní hodnoty – **kategorie**: **Ano, Ne** **(Přítomnost, Nepřítomnost**, apod.).

# **Testování nezávislosti kvalitativních znaků 2**

Uvažujte soubor se dvěma **kvalitativními** znaky ***A*** a ***B***

Znak ***A*** má *r* možných kategorií hodnot

označených: $A_1, A_2, \ldots, A_r$

znak ***B*** má *s* možných kategorií hodnot:

$$B_1, B_2, \ldots, B_s$$

Výsledek celého složeného experimentu lze shrnout do **kontingenční tabulky**:

# Kontingenční tabulka:

| Kategorie znaku $A/B$ | $B_1$ | $B_2$ | $B_3$ | ................. | $B_s$ | Součet |
|---|---|---|---|---|---|---|
| $A_1$ | $n_{11}$ | $n_{12}$ | $n_{13}$ | .................. | $n_{1s}$ | $n_{1.}$ |
| $A_2$ | $n_{21}$ | $n_{22}$ | $n_{23}$ | .................. | $n_{2s}$ | $n_{2.}$ |
| $A_3$ | $n_{31}$ | $n_{32}$ | $n_{33}$ | .................. | $n_{3s}$ | $n_{3.}$ |
| ........... | ..... | .... | .... | .................. | ....... | ................ |
| $A_r$ | $n_{r1}$ | $n_{r2}$ | $n_{r3}$ | .................. | $n_{rs}$ | $n_{r.}$ |
| Součet | $n_{.1}$ | $n_{.2}$ | $n_{.3}$ | .................. | $n_{.s}$ | $n$ |

# Kontingenční tabulka (čtyřpolní)

Příklad: vzhled vers. hmotnost výrobku

| Vzhled / Hmotnost výrobků | Vyhovující hmotnost | Nevyhovující hmotnost | Součet- Marg. četnost |
|---|---|---|---|
| Vyhovující vzhled | 239 | 60 | 299 |
| Nevyhovující vzhled | 14 | 7 | 21 |
| Součet - Marg. četnost | 253 | 67 | 320 |

# Chi-kvadrát test nezávislosti: Algoritmus 1

**Krok 1. Nulová hypotéza $H_0$:**

Vzhled výrobku nezávisí na hmotnosti
(rozdíly u vzorku jsou pouze dílem náhody).

**Očekávané četnosti:** $E_{11} = 253*299/320 = 236{,}4$

$E_{21} = 253*21/320 = 16{,}6$

$E_{12} = 67*299/320 = 62{,}6$

$E_{22} = 67*21/320 = 4{,}4$

**Pozorované četnosti:** $O_{11} = 239$, $O_{12} = 14$, $O_{21} = 60$, $O_{22} = 7$

**Krok 2. Testové kritérium $X^2$:** $$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 2{,}086$$

$df = (r-1)(s-1)$ počet stupňů volnosti ( $k = (2-1)(2-1) = 1$)

## Očekávané četnosti:

$$\frac{\text{Ocekáv\_č\_}i,j}{\text{celk.c.}} = \frac{\text{Marg\_č\_}i}{\text{celk.č.}} \times \frac{\text{Marg\_č\_}j}{\text{celk.č.}}$$

$$\boxed{\text{Očekáv\_č\_}i,j = \text{Marg\_č\_}i \times \text{Marg\_č\_}j \,/\, \text{celk\_č}}$$

**Příklad:**

$i$ = 1: Hmotnost-Nevyhovující

$j$ = 2 : Vzhled-Vyhovující

celk_č = 320

$E_{12}$ = Očekáv_č_1,2 = 299*67/320 = 62,6

atd.

# Chi-kvadrát test nezávislosti: Algoritmus 2

**Krok 3.** Porovnání hodnoty vypočítaného kritéria s tabulkovou kritickou hodnotou rozdělení

kde $\alpha = 0{,}10$ je zadaná hladina významnosti.

V každé kategorii má být alespoň 5 hodnot!

Jestliže $X^2 = 2{,}1 < \chi^2_{0,1}(1) = 2{,}7$ potom $H_0$ nezamítáme!

**Alternativně:**

Pro hodnotu $X^2$ zjistíme $p$-hodnotu (tj. signifikanci - - má být menší než 0,1)

$p = \text{CHIDIST}(2{,}1;1) = 0{,}147$ - tedy $H_0$ nezamítáme!

## **Čtyřpolní tabulka** – kontingenční tabulka 2 x 2:

| Znak2 | | | **Součet** |
|---|---|---|---|
| Znak1 | $h_1$ | $h_2$ | |
| $h_1$ | $A$ | $B$ | $A+B$ |
| $h_2$ | $C$ | $D$ | $C+D$ |
| **Součet** | $A+C$ | $B+D$ | $n$ |

**Kritérium:**

$$X^2 = \frac{n(AD - BC)^2}{(A+B)(C+D)(A+C)(B+D)}$$

**Jestliže** $X^2 > \chi^2_\alpha(1)$

**pak $H_0$ zamítáme, jinak ji nezamítáme!**

# Příklad: Vzhled vers. Hmotnost

A = 239, B = 60, C = 14, D = 7

$$X^2 > \chi^2_{0,1}(1) = 2,7$$

$$X^2 = \frac{n(AD - BC)^2}{(A+B)(C+D)(A+C)(B+D)} = 2,1$$

## **Příklad 4** – Vliv kouření na úmrtnost v Karviné

Kontingenční tabulka pro 2917 zemřelých v Karviné v roce 1998

*Kouření* versus *Počet zemřelých na rakovinu plic*

| Pozorované čet. | zemřel RP | zemřel JINAK |
|---|---|---|
| kouření ANO | 137 | 817 |
| kouření NE | 198 | 1765 |

Analyzujte, zda kouření respondentů ovlivnilo úmrtnost na rakovinu plic (*RP*)

Použijte Chi-kvadrát test

# Řešení příkladu 4 pomocí Excelu:

| Pozorované čet. | zemřel RP | zemřel JINAK | Suma |
|---|---|---|---|
| kouření ANO | 137 | 817 | 954 |
| kouření NE | 198 | 1765 | 1963 |
| Suma | 335 | 2582 | 2917 |

| Očekávané čet. | zemřel RP | zemřel JINAK | (Eij-Oij)^2/Eij | |
|---|---|---|---|---|
| kouření ANO | 109,56 | 844,44 | 6,87 | 0,89 |
| kouření NE | 225,44 | 1737,56 | 3,34 | 0,43 |

součet

| | |
|---|---|
| CHI-SQUARE | 11,54 |
| alfa | 0,05 |
| df | 1 |
| CHIINV | 3,8415 |
| CHIDIST=Sig | 0,0007 |

Nulovou hypotézu o nezávislosti znaků zamítáme!

(Úmrtnost na rakovinu plic závisí na kouření respondentů)

$$X^2 = \frac{n(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} = 11{,}54$$