Q2: N=100.000 terms

$$t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}}$$

a) "adolf" apare de 150 ori

"hitler" apare de 200 ori

"adolf hitler" apare de 175 ori

Ho: "adolf" și "hitler" apar împreună întâmplator (independent) =>

$$P(\text{"adolf hitler"}) = P(\text{"adolf"}) * P(\text{"hitler"})$$

$$= \frac{15\cancel{0}}{1000\cancel{0}\cancel{0}} \cdot \frac{2\cancel{0}\cancel{0}}{100.0\cancel{0}\cancel{0}} = \frac{3\cancel{0}}{10.000.00\cancel{0}} \simeq 3 \cdot 10^{-6}$$

expected mean: $\mu = 3 \cdot 10^{-6}$

observed variation: $\sigma^2 = p(1-p) \approx p = s^2 = \bar{x}(1-\bar{x}) \simeq \bar{x} = 175 \cdot 10^{-5}$

observed mean: $\bar{x} = \frac{175}{100.000} = 175 \cdot 10^{-5}$

$$t = \frac{175 \cdot 10^{-5} - 3 \cdot 10^{-6}}{\sqrt{\frac{175 \cdot 10^{-5}}{10^5}}} = \frac{174,7 \cdot 10^{-5}}{\sqrt{\frac{175}{10^{10}}}} = \frac{174,7}{\cancel{10^5}} \cdot \frac{\cancel{10^5}}{13,228} \simeq 13,21$$

Deoarece $t = 13,21 > 2,573$ (critical value) => Ipoteza Ho este falsa => avem de-a face cu o colocatie.

b) "hitler" apare de 200 ori

"industrial" apare de 700 ori

"hitler industrial" apare de 4 ori

Ho: "hitler" și "industrial" apar împreună independent =>

$$P(\text{"hitler industrial"}) = P(\text{"hitler"}) \cdot P(\text{"industrial"})$$

$$= \frac{2\cancel{0}\cancel{0}}{1000\cancel{0}\cancel{0}} \cdot \frac{7\cancel{0}\cancel{0}}{100.0\cancel{0}\cancel{0}} = \frac{14}{1.000.000} = 14 \cdot 10^{-6}$$

$\mu = 14\cdot10^{-6}$

$s^2 = \bar{x}(1-\bar{x}) \approx \bar{x} = 4\cdot10^{-5}$

$\bar{x} = 4\cdot10^{-5}$

$$\Rightarrow t = \frac{4\cdot10^{-5} - 14\cdot10^{-6}}{\sqrt{\frac{4\cdot10^{-5}}{10^{5}}}} = \frac{(4-1,4)\cdot10^{-5}}{\sqrt{\frac{4}{10^{10}}}} = \frac{2,6}{\cancel{10^{5}}}\cdot\frac{\cancel{10^{5}}}{2} = 1,3$$

Deoarece $t = 1,3 < 2,573$ (the critical value) $\Rightarrow$ Ipoteza Ho este adevărată $\Rightarrow$ "hitler industrial" nu este o colocație.

c) "hitler" → 200 apariții

"revolution" → 900 apariții

"hitler revolution" → 14 apariții

H_0: "hitler" și "revolution" apar împreună întâmplător $\Rightarrow$

P("hitler revolution") = P("hitler")·P("revolution")

$$= \frac{2\cancel{00}}{100.0\cancel{00}}\cdot\frac{9\cancel{00}}{100.0\cancel{00}} = 18\cdot10^{-6}$$

$\mu = 18\cdot10^{-6}$

$\bar{x} = 14\cdot10^{-5}$

$s^2 = \bar{x}(1-\bar{x}) \simeq \bar{x} = 14\cdot10^{-5}$

$$\Rightarrow t = \frac{14\cdot10^{-5} - 18\cdot10^{-6}}{\sqrt{\frac{14\cdot10^{-5}}{10^{5}}}} = \frac{12,2}{\cancel{10^{5}}}\cdot\frac{\cancel{10^{5}}}{3,741} = 3,261$$

Deoarece $t = 3,261 > 2,573$ (the critical value) $\Rightarrow$ Ipoteza este falsă $\Rightarrow$ "hitler revolution" este o colocație.

d) "revolution" → 900 apariții

"hitler" → 200 apariții

"revolution hitler" → 25 apariții

H_0: "revolution" și "hitler" apar împreună în mod independent =>

$$P(\text{"revolution hitler"}) = P(\text{"revolution"}) \cdot P(\text{"hitler"})$$
$$= \frac{9\cancel{0}\cancel{0}}{100.0\cancel{0}\cancel{0}} \cdot \frac{2\cancel{0}\cancel{0}}{100.0\cancel{0}\cancel{0}} = 18 \cdot 10^{-6}$$

$\mu = 18 \cdot 10^{-6}$

$\overline{X} = 25 \cdot 10^{-5}$

$s^2 \simeq \overline{X} = 25 \cdot 10^{-5}$

$$\Rightarrow t = \frac{25 \cdot 10^{-5} - 18 \cdot 10^{-6}}{\sqrt{\frac{25 \cdot 10^{-5}}{10^5}}} = \frac{23,2}{\cancel{10^5}} \cdot \frac{\cancel{10^5}}{5} = 4,64$$

Deoarece $t = 4,64 > 2,573$ (critical value) => Ip. H_0 este falsă =>
"revolution hitler" este o colocație.

e) "industrial" → 700 apariții
"revolution" → 900 apariții
"industrial revolution" → 250 apariții

H_0: "industrial" și "revolution" apar împreună întâmplător =>

$$P(\text{"industrial revolution"}) = P(\text{"industrial"}) \cdot P(\text{"revolution"})$$
$$= \frac{7\cancel{0}\cancel{0}}{100.0\cancel{0}\cancel{0}} \cdot \frac{9\cancel{0}\cancel{0}}{100.0\cancel{0}\cancel{0}} = 63 \cdot 10^{-6}$$

$\mu = 63 \cdot 10^{-6}$

$\overline{X} = 250 \cdot 10^{-5}$

$s^2 \simeq \overline{X} = 250 \cdot 10^{-5}$

$$\Rightarrow t = \frac{250 \cdot 10^{-5} - 63 \cdot 10^{-6}}{\sqrt{\frac{250 \cdot 10^{-5}}{10^5}}} = \frac{243,7}{\cancel{10^5}} \cdot \frac{\cancel{10^5}}{15,811} = 15,413$$

Deoarece $t = 15,413 > 2,573$ (the critical value) => Ip. H_0 este falsă =>
"industrial revolution" este colocație.

Q4:

	w_1 = garden	$w_1 \neq$ garden
w_2 = soil	15	50
$w_2 \neq$ soil	200	400

	w_1 = watch	$w_1 \neq$ watch
w_2 = dog	20	50
$w_2 \neq$ dog	200	1000

a) Chi-squared for "garden soil"

$$\chi_1^2 = \frac{(15+50+200+400)(15\cdot 400 - 200\cdot 50)^2}{(15+200)(50+400)(15+50)(200+400)}$$

$$= \frac{665\cdot(6000-10.000)^2}{215\cdot 450\cdot 65\cdot 600}$$

$$= \frac{665\cdot 16.000.\cancel{000}}{215\cdot 45\cancel{0}\cdot 65\cdot 6\cancel{00}} = \frac{10.640.00\cancel{0}}{3.773.25\cancel{0}} = 2,819$$

b) Chi-squared for "watch dog"

$$\chi_2^2 = \frac{(20+50+200+1000)(20\cdot 1000 - 200\cdot 50)^2}{(20+200)\cdot(50+1000)\cdot(20+50)(200+1000)} =$$

$$= \frac{1270\cdot(10^4)^2}{220\cdot 1050\cdot 70\cdot 1200} =$$

$$= \frac{1270\cdot 1000\cancel{00000}}{22\cancel{0}\cdot 105\cancel{0}\cdot 7\cancel{0}\cdot 12\cancel{00}} = \frac{1.270.00\cancel{0}}{194.04\cancel{0}} = 6,545$$

c) degree of freedom: $r = c = 2 \Rightarrow (r-1)(c-1) = 1$
$p = 0,05$ $\Big\}$ $\Rightarrow$ critical value $= 3,84$
(valoare luată din tabelul de la χ^2)

Deoarece $\chi_1^2 = 2,819 < 3,84$ (the critical value) $\Rightarrow$ "garden soil" nu este o colocație, iar pt. că $\chi_2^2 > \chi_1^2 \Rightarrow$ "watch dog" are șanse mai mari decât "garden soil" să fie o colocație.

Deoarece $\chi_2^2 = 6,545 > 3,84$ (critical value) $\Rightarrow$ "watch dog" chiar este o colocație.