13. cvičení z STP

13. - 17. května 2019

Příklad 11.2.

Příklad 13.1 *U* 64 *praktických lékařů byl naměřen výběrový průměr počtu pacientů za den* 23*, výběrový rozptyl pak byl roven* 36*, rozdělení počtu pacientů není známé.*

(a) Sestrojte (asymptotický) oboustranný interval pro střední hodnotu počtu pacientů o spolehlivosti 95%*.*

(b) Otestujte na hladině 5%*, zda skutečná střední hodnota počtu pacientů za den může být považována za rovnu* 25*.*

Řešení:

Máme veličiny

$$X_i = \text{“počet pacientů u } i\text{-tého lékaře za den”}$$

pro $i = 1, \ldots, n$, kde $n = 64$, které budeme pokládat za nezávislé. Jejich rozdělení není známé (i když bychom asi očekávali, že by mohlo být Poissonovo). Vzhledem k této neznalosti budeme potřebovat pracovat s asymptotickým přiblížením.

(a) Asymptotický oboustranný interval pro střední hodnotu o spolehlivosti $1 - \alpha$ se bude podobat tomu, který se odvozuje, pokud X_i mají normální rozdělení. Rozdíl bude jen v tom, že kvantil $t_{1-\frac{\alpha}{2};\, n-1}$ pro Studentovo rozdělení s $n-1$ stupni volnosti (který používáme, když X_i mají normální rozdělení) se nahradí jeho asymptotickou hodnotou když $n \to \infty$, která odpovídá kvantilu $u_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ pro rozdělení $N(0,1)$.

Příslušný asymptotický interval o spolehlivosti $1 - \alpha = 95\%$ tedy je:

$$\langle \mu_L,\ \mu_U \rangle := \left\langle\ \overline{\boldsymbol{x}} - \frac{s_{\boldsymbol{x}}}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}} \quad , \quad \overline{\boldsymbol{x}} + \frac{s_{\boldsymbol{x}}}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}}\ \right\rangle$$

kde $\overline{\boldsymbol{x}} = 23$ je výběrový průměr a $s^2_{\boldsymbol{x}} = 36$ je výběrový rozptyl. Pro $\alpha = 5\%$ máme hodnotu kvantilu $u_{1-\frac{\alpha}{2}} = u_{0.975} = \Phi^{-1}(0.975) \doteq 1.96$. Po dosazení máme tedy

$$\langle \mu_L,\ \mu_U \rangle := \left\langle\ 23 - \frac{\sqrt{36}}{\sqrt{64}} \cdot 1.96 \quad , \quad 23 + \frac{\sqrt{36}}{\sqrt{64}} \cdot 1.96\ \right\rangle \doteq \langle\ 22.28,\ 23.72\ \rangle$$

Tento interval se pochopitelně MĚNÍ s každým měřením (protože je závislý na naměřených vstupech), a jeho smysl je ten, že skutečná hodnota $\mu = E(X)$ (která se NEMĚNÍ!) bude obsažena v tomto (obecně proměnném intervalu) s pravděpodobností 95%.

(b) Podle zadání máme na hladině $\alpha = 5\%$ otestovat hypotézu o střední hodnotě $\mu = E(X)$ tvaru

$$\mathbf{H}_0 : \mu = \mu_0$$

proti alternativní hypotéze:

$$\mathbf{H}_A : \mu \neq \mu_0\ .$$

kde $\mu_0 = 25$.

Pomocí intervalového odhadu:

Využijeme už spočítaného asymptotického oboustranného intervalu $\langle \mu_L,\ \mu_U \rangle$ pro střední hodnotu o spolehlivosti 95%. Podle toho, co jsme uvedli výše v části (a), je pravděpodobnost, že střední hodnota $\mu = E(X)$ bude obsažena v (proměnném) intervalu $\langle \mu_L,\ \mu_U \rangle$, rovna 95%. Tedy mimo tento interval se ocitne jen v 5% případů.

Jestliže předpokládáme, že $\mu = \mu_0$ (tj. hypotézu $\mathbf{H}_0$), bude kritérium pro její zamítnutí na hladině α přirozeně tvaru:

$$\mu_0 \notin \langle \mu_L,\ \mu_U \rangle \quad \Leftrightarrow \quad \text{zamítáme } \mathbf{H}_0 \text{ (na hladině } \alpha) \ .$$

A protože skutečně nakonec máme, že $\mu_0 = 25 \notin \langle\ 22.28,\ 23.72\ \rangle = \langle \mu_L,\ \mu_U \rangle$, tak hypotézu $\mathbf{H}_0$ **ZAMÍTÁME** na hladině 5%.

Pomocí testovací statistiky:

Podmínka pro zamítnutí $\mathbf{H}_0$ na hladině α se dá z formy pro interval spolehlivosti

$$\mu_0 \notin \left\langle\ \overline{\boldsymbol{x}} - \frac{s_{\boldsymbol{x}}}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}} \quad , \quad \overline{\boldsymbol{x}} + \frac{s_{\boldsymbol{x}}}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}}\ \right\rangle$$

ekvivalentně přepsat jako

$$|t| > u_{1-\frac{\alpha}{2}}$$

kde $t = \frac{\overline{\boldsymbol{x}} - \mu_0}{s_{\boldsymbol{x}}}\sqrt{n}$, což je (podobně jako v **Příkladu 11.3**) hodnota testovací veličiny (tzv. *statistiky*):

$$T = \frac{\overline{\boldsymbol{X}} - \mu_0}{S_{\boldsymbol{X}}}\sqrt{n}$$

Protože však neznáme rozdělení veličin X_i, neznáme ani *přesné* rozdělení této statistiky T. To ale na druhou stranu nevadí, protože pro dost velká n nakonec bude mít veličina T přibližné rozdělení $N(0,1)$ (bez ohledu na počáteční rozdělení rozdělení veličin X_i). To je tedy důvod, proč se pak v zamítacím kritériu objevují kvantily pro norm. rozdělení. Co přesně znamená "dost velká n", závisí pochopitelně na tom, jak "divoké" je rozdělení veličin X_i. Když si teď připustíme, že X_i by mohly mít skutečně Poissonovo rozdělení, tak i relativně malé hodnoty n (my máme $n = 64$) mohou být dostatečné pro použití asymptotiky.

Shrňme si to tedy tak, ze kritérium pro zamítnutí $\mathbf{H}_0$ (na hladině α) je tvaru

$$|t| > u_{1-\frac{\alpha}{2}} \quad \Leftrightarrow \quad \text{zamítáme } \mathbf{H}_0 \text{ (na hladině } \alpha) \ .$$

Při konkretním dosazení máme

$$t \ = \ \frac{\overline{\boldsymbol{x}} - \mu_0}{s_{\boldsymbol{x}}}\sqrt{n} \ = \ \frac{23 - 25}{\sqrt{36}}\sqrt{64} \ = \ -\frac{16}{3} \ \doteq \ -5.33$$

a tudíž

$$|t| \doteq |5.33| > 1.96 \doteq u_{0.975}$$

což znamená, že hypotézu $\mathbf{H}_0$ (opět) **ZAMÍTÁME** na hladině 5%.

(Výsledek musel samozřejmě dopadnout stejně jako pomoci intervalu spolehlivosti, protože je to ekvivalentní princip.)

Poznámky k testu dobré shody: Chceme otestovat (na hladině α), jestli daná veličina X s konečně mnoha (navzájem různými!) hodnotami $a_1, \ldots, a_k$ (ne nutně číselnými) má předepsané pravděpodobnosti $(p_1, \ldots, p_k)$, tedy nulovou hypotézu

$$\mathbf{H}_0 : P(X = a_i) = p_i \text{ pro všechna } i \in \{1, \ldots, k\}$$

proti alternativní hypotéze:

$$\mathbf{H}_A : P(X = a_{i_0}) \neq p_{i_0}, \text{ pro alespoň jedno } i_0 \in \{1, \ldots, k\}$$

Při n pokusech s veličinou X si pro $i = 1, \ldots, k$ označme veličiny

$$N_i = \text{“počet výskytů případu } X = a_i \text{ při } n \text{ pokusech”} .$$

Máme tedy náhodný vektor

$$\boldsymbol{N} = (N_1, \ldots, N_k)$$

a vztah $N_1 + \cdots + N_k = n$. Už z toho vidíme, že veličiny N_i nejsou nezávislé, ale zase k té nezávislosti tak daleko nemají. Náhodný vektor $\boldsymbol{N}$ má tzv. multinomické rozdělení a jednotlivá marginální rozdělení veličin jsou binomická, konkrétně $N_i \sim \mathrm{Bi}(n, p_i)$. Speciálně tedy $E(N_i) = n \cdot p_i$.

Jako testovací veličinu zde používáme:

$$T = \sum_{i=1}^{k} \frac{(N_i - n \cdot p_i)^2}{n \cdot p_i}$$

která má asymptoticky (tj. pro $n \to \infty$) tzv. χ^2-rozdělení s $k - 1$ stupni volnosti. Pro praktické použití této asymptotiky se obvykle požaduje, aby platilo, že

$$n \cdot p_i \geq 5 \quad \text{pro všechna} \quad i \in \{1, \ldots, k\} .$$

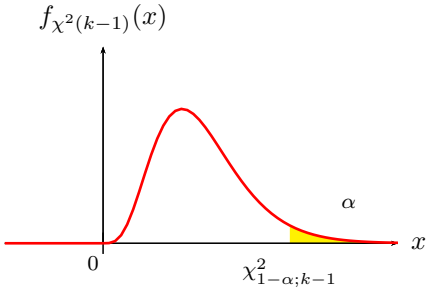
Hodnoty $n \cdot p_i$ se označují jako tzv. *teoretické četnosti*.

Pokud tedy platí nulová hypotéza $\mathbf{H}_0$, měly by být hodnoty veličiny T malé. Jestliže hodnoty T budou příliš velké, bude to důvod k zamítnutí nulové hypotézy.

Jak určit hranici, kde už nastane zamítnutí: veličina T má (přibližně) $\chi^2_{(k-1)}$ rozdělení, tedy platí

$$P_{(H_0 \text{ platí})}\left(T > \chi^2_{1-\alpha;\, k-1}\right) \doteq \alpha$$

kde $\chi^2_{1-\alpha;\, k-1}$ je hodnota kvantilu pro $\chi^2_{(k-1)}$ rozdělení (viz obrázek, kde α je velikost žluté plochy pod hustotou $f_{\chi^2(k-1)}(x)$ pro $\chi^2_{(k-1)}$ rozdělení).



Kritérium pro **ZAMÍTNUTÍ** $\mathbf{H}_0$ (na hladině α) proto volíme jako

$$t > \chi^2_{1-\alpha;\, k-1} \quad \Leftrightarrow \quad \text{zamítáme } \mathbf{H}_0 \text{ (na hladině } \alpha) .$$

Z definice chyby 1. druhu, tj.

$$\text{nastává chyba 1. druhu} \quad \Leftrightarrow \quad (\text{hypotéza } \mathbf{H}_0 \text{ platí} \quad \& \quad \text{my ji zamítáme})$$

pak totiž máme, že

$$P_{(H_0 \text{ platí})}\Big(\text{nastává chyba 1. druhu}\Big) = P_{(H_0 \text{ platí})}\Big(\text{zamítáme } \mathbf{H}_0 \text{ (na hladině } \alpha)\Big) =$$

$$= P_{(H_0 \text{ platí})}\big(T > \chi^2_{1-\alpha;\, k-1}\big) \doteq \alpha$$

neboli pravděpodobnost chyby 1. druhu (ovšem za předpokladu platnosti $\mathbf{H}_0$!) je pak omezena hodnotou α.

Příklad 13.2 *Firma má* 3 *pobočky. Dva roky bylo sledováno, která z nich zaznamenala nejvyšší měsíční výnos. Bylo zjištěno, že nejvýnosnější byla první pobočka* $10\times$*, druhá* $6\times$ *a třetí* $8\times$*.*
Je možné říct, že první pobočka je nejvýnosnější $2\times$ *častěji než každá ze zbylých dvou? Testujte na hladině* 5%.

Řešení:
Máme tedy veličinu

$$X = \text{“číslo pobočky, která je zrovna (tj. v daném týdnu) nejvýnosnější”}$$

s $k = 3$ hodnotami $\{\textit{první, druhá, třetí}\}$.

Nejdříve si potřebujeme zjistit, jaké rozdělení

$$P(X = \textit{první}) = p_1, \quad P(X = \textit{druhá}) = p_2, \quad P(X = \textit{třetí}) = p_3$$

vlastně předpokládáme. Z požadavku máme

$$p_1 = 2 \cdot p_2, \quad p_1 = 2 \cdot p_3, \quad p_1 + p_2 + p_3 = 1$$

z čehož dostáváme

$$(p_1, p_2, p_3) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$$

Naše hypotéza tedy je

$\mathbf{H}_0$: veličina X má rozdělení s pravděpodobnostmi $(p_1, p_2, p_3) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$,

a alternativní hypotéza bude:

$\mathbf{H}_A$: veličina X má rozdělení *jiné* než $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$.

Využijeme test dobré shody. Celkový počet měření (tj. počet týdnů) je $n = 10 + 6 + 8 = 24$. Pro přehlednost si vypíšeme tabulku s jednotlivými četnostmi (pozorovanými i teoretickými):

i (pobočky)	první	druhá	třetí
n_i (pozorované četnosti)	10	8	6
p_i (teoretické pravděpodobnosti)	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
$n \cdot p_i$ (teoretické četnosti)	$24 \cdot \frac{1}{2} = 12$	$24 \cdot \frac{1}{4} = 6$	$24 \cdot \frac{1}{4} = 6$

Vidíme, že všechny teoretické četnosti jsou ≥ 5, takže skutečně můžeme použít asymptotické přiblížení pro testovací statistiku T (ta tedy bude mít χ^2-rozdělení). Teď už si jen spočítáme hodnotu této statistiky

$$t = \sum_{i=1}^{3} \frac{(n_i - np_i)^2}{np_i} = \frac{(10-12)^2}{12} + \frac{(8-6)^2}{6} + \frac{(6-6)^2}{6} = \frac{1}{3} + \frac{2}{3} + 0 = 1$$

a porovnáme s kvantilem χ^2-rozdělení s $k - 1 = 3 - 1 = 2$ stupni volnosti:

$$t = 1 \not> 5.99 \doteq \chi^2_{0.95;\, 2} = \chi^2_{1-\alpha;\, k-1}$$

Protože zamítací kritérium NENÍ splněno, tak $\mathbf{H}_0$ **NEZAMÍTÁME** (na hladině α).

Poznámky k testu nezávislosti: Máme veličiny

- X s (různými) hodnotami $\{a_1, \dots, a_k\}$ a
- Y s (různými) hodnotami $\{b_1, \dots, b_\ell\}$

a chceme otestovat (na hladině α), hypotézu

$\mathbf{H}_0$: rozdělení veličin X a Y jsou *nezávislá*

proti alternativní hypotéze:

$\mathbf{H}_A$: rozdělení veličin X a Y jsou *závislá*

Při n pokusech s náhodným vektorem (X, Y) si pro $i = 1, \dots, k$ označme veličiny

$$N_{i,j} = \text{“počet výskytů případu } (X, Y) = (a_i, b_j) \text{ při } n \text{ pokusech”} .$$

a opět máme náhodný vektor

$$\boldsymbol{N} = (N_{1,1}, \dots, N_{k,\ell})$$

s multinomickým rozdělením. Marginální rozdělení jednotlivých veličin $N_{i,j}$ jsou opět binomická a *za předpokladu nezávislosti X a Y* mají střední hodnotu

$$E(N_{i,j}) = n \cdot P(X = a_i, Y = b_j) \overset{\text{(nezáv.)}}{=} n \cdot P(X = a_i) \cdot P(Y = b_j) .$$

Podobně jako v testu dobré shody by tyto střední hodnoty představovaly teoretické četnosti až na to, že pravděpodobnosti $P(X = a_i)$ a $P(Y = b_j)$ nemáme v hypotéze uvedeny. Proto je odhadneme jako

$$P(X = a_i) \doteq \frac{n_{i,\bullet}}{n} \quad \text{a} \quad P(Y = b_j) \doteq \frac{n_{\bullet,j}}{n}$$

kde $n_{i,\bullet}$ a $n_{\bullet,j}$ jsou naměřené hodnoty veličin

$$N_{i,\bullet} = \sum_{j=1}^{\ell} N_{i,j} = \text{“počet výskytů případu } X = a_i \text{ při } n \text{ pokusech”}$$

$$N_{\bullet,j} = \sum_{i=1}^{k} N_{i,j} = \text{“počet výskytů případu } Y = b_j \text{ při } n \text{ pokusech”}$$

což jsou tzv. *marginální četnosti*.

Z tohoto důvodu jako testovací veličinu volíme:

$$T = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{\left(N_{i,j} - \frac{N_{i,\bullet} \cdot N_{\bullet,j}}{n}\right)^2}{\frac{N_{i,\bullet} \cdot N_{\bullet,j}}{n}}$$

která má asymptoticky (tj. pro $n \to \infty$) opět χ^2-rozdělení, tentokrát ale s $(k-1) \cdot (\ell - 1)$ stupni volnosti. Pro praktické použití této asymptotiky se obvykle opět požaduje, aby platilo, že

$$\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n} \geq 5 \quad \text{pro všechna} \quad i = 1, \dots, k \text{ a } j = 1, \dots, \ell.$$

Kritérium pro **ZAMÍTNUTÍ** $\mathbf{H}_0$ (na hladině α) volíme podobně jako u testu dobré shody a sice

$$t > \chi^2_{1-\alpha;\,(k-1)\cdot(\ell-1)} \quad \Leftrightarrow \quad \text{zamítáme } \mathbf{H}_0 \text{ (na hladině } \alpha) .$$

Příklad 13.3 *Na $n = 100$ osobách byla pozorována barva očí a vlasů. Naměřeny byly následující sdružené četnosti:*

Oči \ *Vlasy*	*tmavé*	*světlé*
modré	10	20
šedé	10	10
hnědé	40	10

(a) Jsou barvy očí a vlasů nezávislé? Testujte na hladině 5%.

(b) Otestujte na hladině 5%*, jestli je v populaci stejně tmavovlasých jako světlovlasých.*

Řešení:

Označme si veličiny

X= *"barva očí daného člověka"*

Y= *"barva vlasů daného člověka"*

a dál budeme pracovat s náhodným vektorem (X, Y), tj. u daného člověka budeme zjišťovat barvu očí a barvu vlasů.

(a) Budeme testovat hypotézu:

$\mathbf{H}_0$: rozdělení veličin X a Y jsou *nezávislá*

proti alternativní hypotéze:

$\mathbf{H}_1$: rozdělení veličin X a Y jsou *závislá*.

na hladině významnosti $\alpha = 5\%$.

Četnost případu $(X, Y) = (i, j)$ v tabulce označme jako $n_{i,j}$ a marginální četnosti pak budou

$n_{i,\bullet} = \sum_j n_{i,j}$ pro případ $X = i$

$n_{\bullet,j} = \sum_i n_{i,j}$ pro případ $Y = j$.

což jsou součty v řádcích a sloupcích tabulky:

$n_{i,j}$ $(Y=)\ j$ / $(X=)\ i$	tmavé	světlé	$n_{i,\bullet}$
modré	10	20	30
šedé	10	10	20
hnědé	40	10	50
$n_{\bullet,j}$	60	40	

Za předpokladu $\mathbf{H}_0$ pak jako teoretické četnosti budeme chápat hodnoty $\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}$ v této tabulce:

$\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}$ $(Y=)\ j$ / $(X=)\ i$	tmavé	světlé	$n_{i,\bullet}$
modré	$\frac{30 \cdot 60}{100} = 18$	$\frac{30 \cdot 40}{100} = 12$	30
šedé	$\frac{20 \cdot 60}{100} = 12$	$\frac{20 \cdot 40}{100} = 8$	20
hnědé	$\frac{50 \cdot 60}{100} = 30$	$\frac{50 \cdot 40}{100} = 20$	50
$n_{\bullet,j}$	60	40	

Podmínka na tyto teoretické (tj. očekávané) četnosti ≥ 5 je splněna, takže test nezávislosti můžeme použít. Pro hodnotu testovací statistiky dostaneme

$$t = \sum_{i,j} \frac{\left(n_{i,j} - \frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}\right)^2}{\frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}} =$$

$$= \frac{(10-18)^2}{18} + \frac{(10-12)^2}{12} + \frac{(40-30)^2}{30} + \frac{(20-12)^2}{12} + \frac{(10-8)^2}{8} + \frac{(10-20)^2}{20} = 18 + \frac{1}{18} \doteq 18.056 \ .$$

Tuto hodnotu dále porovnáme s hodnotou kvantilu χ^2 pro $(k-1)(\ell-1)$ stupňů volnosti, kde k je počet položek veličiny X a ℓ je počet položek veličiny Y. Tento počet je nyní jiný, než by byl u "obvyklého" testu dobré shody s $k\cdot\ell$ položkami, protože data jsme použili k odhadu marginálních pravděpodobností.

Kritérium pro **ZAMÍTNUTÍ $\mathbf{H}_0$** (na hladině α) bude tedy tvaru

$$t > \chi^2_{1-\alpha;\ (k-1)(\ell-1)} \quad \Leftrightarrow \quad \text{zamítáme } \mathbf{H}_0 \text{ (na hladině } \alpha) \ .$$

Hledaný kvantil je

$$\chi^2_{1-\alpha;\ (3-1)(2-1)} = \chi^2_{0.95;\ 2} \doteq 5.992 \ .$$

Protože

$$t \doteq 18.056 > 5.992 \doteq \chi^2_{0.95;\ 2} \ ,$$

hypotézu o nezávislosti **ZAMÍTÁME**.

- -

(b) V tomto případě budeme uvažovat pouze veličinu Y a testovat (na hladině $\alpha = 5\%$) hypotézu

$\widetilde{\mathbf{H}}_0$: veličina Y má rozdělení s pravděpodobnostmi $(p_1, p_2) = \left(\frac{1}{2}, \frac{1}{2}\right)$,

proti alternativní hypotéze

$\widetilde{\mathbf{H}}_A$: veličina Y má rozdělení *jiné* než $\left(\frac{1}{2}, \frac{1}{2}\right)$.

Využijeme teď opět test dobré shody. Celkový počet měření je zase $n = 100$. Naměřené četnosti odpovídají už spočítaným marginálním četnostem pro hodnoty veličiny Y, tedy $n_i = n_{\bullet,i}$. Pro přehlednost si zase vypíšeme tabulku s jednotlivými četnostmi (pozorovanými i teoretickými):

i (barvy vlasů)	tmavé	světlé
n_i (pozorované četnosti)	60	40
p_i (teoretické pravděpodobnosti)	$\frac{1}{2}$	$\frac{1}{2}$
$n \cdot p_i$ (teoretické četnosti)	$100 \cdot \frac{1}{2} = 50$	$100 \cdot \frac{1}{2} = 50$

Vidíme, že všechny teoretické četnosti jsou ≥ 5, takže můžeme použít asymptotické přiblížení pro testovací statistiku T. Teď už si jen spočítáme hodnotu této statistiky

$$t = \sum_{i=1}^{2} \frac{(n_i - np_i)^2}{np_i} = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} = 2 + 2 = 4$$

a porovnáme s kvantilem χ^2-rozdělení s $k - 1 = 2 - 1 = 1$ stupněm volnosti:

$$t = 4 > 3.84 \doteq \chi^2_{0.95;\ 1} = \chi^2_{1-\alpha;\ k-1}$$

Protože zamítací kritérium JE splněno, tak $\widetilde{\mathbf{H}}_0$ **ZAMÍTÁME** (na hladině α).