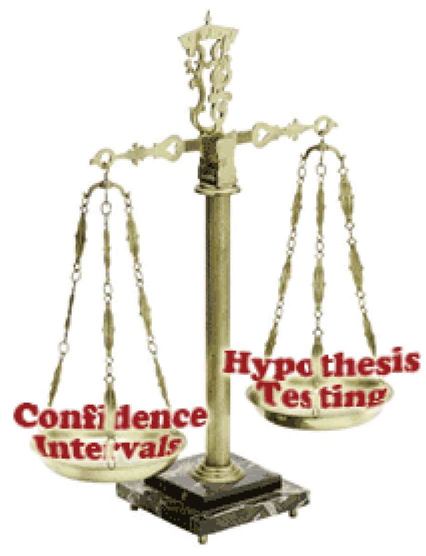
ÚVOD
DO TESTOVÁNÍ HYPOTÉZ

Martina Litschmannová

Základní metody statistické indukce

- Intervalové odhady (angl. confidence intervals) – umožňují odhadnout nejistotu v odhadu parametru náhodné veličiny
- Testování hypotéz (angl. hypothesis testing) - umožňuje posoudit, zda experimentálně získaná data nepopírají předpoklad, který jsme **před** provedením testování učinili.

Používáme, chceme-li určit velikost parametru NV, resp. velikost efektu (rozdílu, resp. poměru parametrů dvou NV).



Používáme, chceme-li ověřit platnost předem definované hypotézy (s předem danou hladinou významností).

zdroj: http://www.nedarc.org/

Co je to statistická hypotéza?

Statistická hypotéza – předpoklad (tvrzení) o rozdělení náhodné veličiny

Co je zdrojem statistických hypotéz?

- předchozí zkušenosti,
- teorie, kterou je třeba doložit,
- požadavky na kvalitu produktu,
- dohady založené na náhodném pozorování…

Co je to statistická hypotéza?

Statistická hypotéza – předpoklad (tvrzení) o rozdělení náhodné veličiny

Příklady statistických hypotéz:

- Střední životnost žárovek Ed je nižší než výrobcem udávaných 5 let.
- Mortalita je u laparoskopických operací nižší než u operací konvenčních.
- Průměrné výsledky srovnávacích testů závisí na typu absolvované střední školy.
- Pořízený datový soubor je výběrem z populace mající normální rozdělení.

Poznámka: Rozdíl (resp. poměr) parametru náhodné veličiny a jeho očekávané hodnoty, popřípadě rozdíl (resp. poměr) parametrů náhodných veličin nazýváme efekt.

Jaké typy statistických hypotéz rozlišujeme?

- Parametrická statistická hypotéza – tvrzení ohledně efektu
  - Hypotézy o parametru jedné populace (o střední hodnotě, rozptylu, mediánu, parametru binomického rozdělení, ...)
  - Hypotézy o parametrech dvou populací (srovnávací testy)
  - Hypotézy o parametrech více než dvou populací (ANOVA, Kruskalův-Wallisův test, ...)

- Neparametrická statistická hypotéza – tvrzení o jiné vlastnosti rozdělení náhodné veličiny než o jejím parametru (např. hypotézy o typu rozdělení NV, hypotézy o závislosti NV, ...)

Jak ověřit, zda je statistická hypotéza pravdivá?

Příklad: Domníváme se, že střední hodnota obsahu cholesterolu v krvi je u české populace 4,7 mmol/l.

$$H_0\colon \mu = 4{,}7$$
$$H_A\colon \mu \neq 4{,}7$$

Jak tento předpoklad ověřit?

- Zjistíme údaje o obsahu cholesterolu v krvi u 100 náhodně vybraných Čechů.
- Průměrný obsah cholesterolu v krvi probandů (tj. jedinců, kteří jsou předmětem zkoumání) byl 5,4 mmol/l.

Jsou tyto výsledky v souladu s naší hypotézou?

- I kdyby byla testovaná hypotéza pravdivá, nelze očekávat, že průměrná hodnota pozorovaná ve výběru bude přesně 4,7 mmol/l.
- **Nulovou hypotézu zamítneme, pokud získané uspořádání výběru bude za předpokladu platnosti nulové hypotézy velmi nepravděpodobné.**

Jak ověřit, zda je statistická hypotéza pravdivá?

Pravdivost nulové hypotézy nelze na základě dat dokázat!!!
Pravdivost nulové hypotézy lze na základě dat pouze vyvrátit.

Nulová hypotéza
(obžalovaný je nevinen)

Alternativní hypotéza
(obžalovaný je vinen)

Data (výběrový soubor)
(svědci)

Testové kritérium
(soudce)

Princip presumpce neviny

Neodsoudí-li soudce obžalovaného, nemusí to znamenat, že je obžalovaný nevinný.
Může to znamenat, že neexistuje dostatek důkazů pro jeho odsouzení!

Co je to testování hypotéz?

Egon Sharpe Pearson (1895-1980)
zdroj: http://www-groups.dcs.st-and.ac.uk

Jerzy Neymann (1894-1981)
zdroj: http://www-history.mcs.st-and.ac.uk/

- Rozhodovací proces, v němž proti sobě stojí nulová a alternativní hypotéza.
- Nulová hypotéza H_0 – tvrzení, že efekt je nulový, resp. že neexistuje závislost, že data mají určitý typ rozdělení, …
- Alternativní hypotéza H_A (H_1) – tvrzení popírající hypotézu nulovou (obvykle to, co chceme dokázat)

Terminologie v praxi (I)

Zadání problému: Ověřte, zda použití bezpečnostních pásů ovlivňuje úmrtnost při dopravních nehodách.

Populace 1 (základní soubor 1): účastníci dopravních nehod, kteří seděli na místech, na nichž je možno používat bezpečnostní pásy a byli připoutáni.

Populace 2 (základní soubor 2): účastníci dopravních nehod, kteří seděli na místech, na nichž je možno používat bezpečnostní pásy a nebyli připoutáni.

Sledovaný statistický znak (náhodná veličina): úmrtnost (relativní četnost zemřelých)

Nulová hypotéza $\boldsymbol{H_0}$: $\pi_A = \pi_N$, kde π_A, resp. π_N označuje úmrtnost účastníků dopravních nehod, kteří byli, resp. nebyli, připoutáni

Alternativní hypotéza $\boldsymbol{H_A}$: $\pi_A \neq \pi_N$ (zadání problému neobsahuje jednostrannou nerovnost.

Terminologie v praxi (II)

Zadání problému*:* Ověřte, zda průměrný plat v ČR je větší než 24 000,- Kč.

Populace (základní soubor): všichni občané ČR pobírající mzdu

Sledovaný statistický znak (náhodná veličina): mzda

Nulová hypotéza $\boldsymbol{H_0}$: $\mu = 24\,000$

Alternativní hypotéza $\boldsymbol{H_A}$: $\mu > 24\,000$

(zadání obsahuje nerovnost v tomto tvaru)

Poznámka: *Průměrný plat zjištěný z výběrového souboru by měl být větší než 2 400,- Kč. Pokud by tomu tak nebylo, měli bychom použít oboustrannou alternativní hypotézu.*

Jak postupovat při testování hypotéz? (klasický přístup)

1. Formulujeme nulovou a alternativní hypotézu.

2. Zvolíme tzv. testovou statistiku, tj. výběrovou charakteristiku, jejíž rozdělení závisí na testovaném parametru θ. (Rozdělení testové statistiky za předpokladu platnosti nulové hypotézy nazýváme nulové rozdělení.)

3. Ověříme předpoklady testu!

4. Určíme kritický obor W^*, tj. množinu, v níž se, za předpokladu platnosti H_0, hodnoty testové statistiky vyskytují s velmi malou pravděpodobností.
   - Doplňkem k W^* je tzv. obor přijetí V^*.
   - Hranici mezi kritickým oborem a oborem přijetí označujeme jako kritická hodnota testu t_{krit}.

5. Na základě konkrétní realizace výběru určíme pozorovanou hodnotu x_{OBS} testové statistiky.

6. Na základě vztahu mezi x_{OBS} a t_{krit} rozhodneme o výsledku testu („Zamítáme H_O.“ nebo „Nezamítáme H_O.“)

Standardním výrobním způsobem lze vyrobit monitory se střední životnosti 1200 hodin a směrodatnou odchylkou 300 hodin. Novou technologií, kterou navrhuje vývojové centrum bylo zkušebně vyrobeno 100 obrazovek, jejichž průměrná životnost byla 1265 hodin. **Jde o kvalitnější technologii, nejde pouze o náhodný rozdíl?**

Řešení:

1. $H_0\!: \ \mu = 1200, H_A\!: \ \mu > 1200$

2. $T(\boldsymbol{X}) = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \to N(0;1)$

testová statistika

nulové rozdělení

Toto platí pouze v případě, že X je náhodný výběr z populace mající normální rozdělení!!!

předpoklady testu

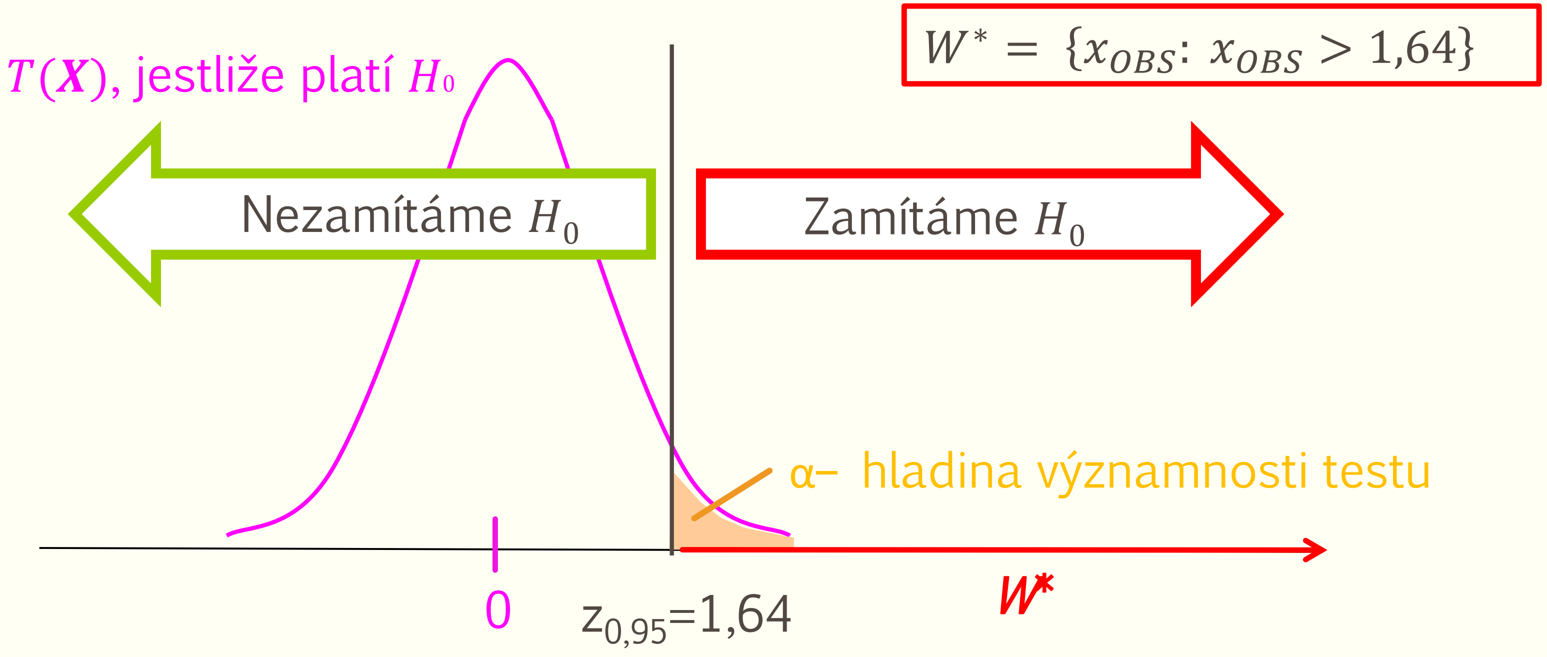
3. Ověření předpokladů testu:

- Zajištění náhodného výběru je důležité již ve fázi plánování experimentu!
- Normalitu výběru je nutno ověřit!!! (Nyní např. pomocí exploračních grafů, později pomocí statistického testu.) V tomto příkladu nemáme k dispozici reálná data. Pokračovat v řešení má smysl pouze tehdy, můžeme-li předpokládat normální rozdělení životnosti obrazovek.

Standardním výrobním způsobem lze vyrobit monitory se střední životnosti 1200 hodin a směrodatnou odchylkou 300 hodin. Novou technologií, kterou navrhuje vývojové centrum bylo zkušebně vyrobeno 100 obrazovek, jejichž průměrná životnost byla 1265 hodin. **Jde o kvalitnější technologii, nejde pouze o náhodný rozdíl?**

Řešení:

4. Pro určení kritického oboru je nutné předem si stanovit, jak „nepravděpodobné" hodnoty testové statistiky již budeme považovat za „velmi nepravděpodobné".



Standardním výrobním způsobem lze vyrobit monitory se střední životnosti 1200 hodin a směrodatnou odchylkou 300 hodin. Novou technologií, kterou navrhuje vývojové centrum bylo zkušebně vyrobeno 100 obrazovek, jejichž průměrná životnost byla 1265 hodin. **Jde o kvalitnější technologii, nejde pouze o náhodný rozdíl?**

Řešení:

4. Pro určení kritického oboru je nutné předem si stanovit, jak „nepravděpodobné“ hodnoty testové statistiky již budeme považovat za „velmi nepravděpodobné“.

$$\boxed{W^* = \{x_{OBS}\colon x_{OBS} > 1{,}64\}}$$

5. $x_{OBS} = T(\boldsymbol{x})|H_0 = \frac{1265-1200}{300}\sqrt{100} = 2{,}17$

6. $x_{OBS} \in W^* \Rightarrow$ Na hladině významnosti 0,05 zamítáme H_0 ve prospěch H_A.
Pozorované zlepšení průměrné životnosti obrazovek je statisticky významné.

Chyba I. a II. druhu

Při testování hypotéz mohou nastat čtyři situace:

hladina významnosti testu

Rozhodnutí

		Nezamítáme H_0	Zamítáme H_0
Skutečnost	Platí H_0	Správné rozhodnutí Pravděpodobnost: $1 - \alpha$	Chyba I. druhu Pravděpodobnost: α
	Platí H_A	Chyba II. druhu Pravděpodobnost: β	Správné rozhodnutí Pravděpodobnost: $1 - \beta$

síla testu

- Jelikož výběr na jehož základě rozhodujeme je náhodný, nelze se chybám I. a II. druhu vyhnout.
- Chtěli bychom mít k dispozici testy s nízkou hladinou významnosti a vysokou sílou testu.

Chyba I. a II. druhu

Sledujte v appletu Chyba I. a II. druhu vztah mezi pravděpodobnostmi těchto chyb.

Závěry:

Chtěli bychom mít k dispozici testy s nízkou hladinou významnosti a vysokou sílou testu - to jsou, bohužel, dva protichůdné požadavky.

- S klesající hladinou významnosti roste pravděpodobnost chyby II. druhu!
- Existuje jediný způsob jak snížit α i β – zvýšení rozsahu výběru.
- Hladinu významnosti α volíme obvykle 0,05 (resp. 0,01).
- Sílu testu lze poté ovlivnit volbou testové statistiky a dostatečného počtu pozorování.

Výšku asijských hybridů lilií lze modelovat **náhodnou veličinou s normálním rozdělením** $N(100; 144)$. Skupina 100 kusů těchto lilií byla pěstována za příznivějších podmínek, aby se zjistilo, zda se výška zvýší.

a) Průměrná výška testovaného vzorku lilií je 102,5 cm. Ověřte klasickým testem, zda lze se spolehlivostí 0,95, resp. 0,99, tvrdit, že nové pěstební podmínky vedly ke zvýšení střední výšky asijských hybridů lilií.

Řešení:

1. $H_0\colon \mu = 100, \quad H_A\colon \mu > 100$

2. $T(\boldsymbol{X}) = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \to N(0;1)$

3. Předpoklady testu nelze popřít.

4. $W^* = \left\{x_{OBS}\colon x_{OBS} > z_{0,95}\right\} = \{x_{OBS}\colon x_{OBS} > 1{,}64\}$

5. $x_{OBS} = T(\boldsymbol{X})|H_0 = \frac{102-100}{12}\sqrt{100} = 1{,}67$

6. $x_{OBS} \in W^* \Rightarrow$ Na hladině významnosti 0,05 zamítáme H_0 ve prospěch H_A. Pozorované zvýšení průměrné výšky asijských hybridů lilií je na hladině významnosti 0,05 statisticky významné.

Výšku asijských hybridů lilií lze modelovat **náhodnou veličinou s normálním rozdělením** $N(100;144)$. Skupina 100 kusů těchto lilií byla pěstována za příznivějších podmínek, aby se zjistilo, zda se výška zvýší.

a) Průměrná výška testovaného vzorku lilií je 102,5 cm. Ověřte klasickým testem, zda lze se spolehlivostí 0,95, resp. 0,99, tvrdit, že nové pěstební podmínky vedly ke zvýšení střední výšky asijských hybridů lilií.

Řešení:

1. $H_0\colon\ \mu = 100,\quad H_A\colon\ \mu > 100$

2. $T(\boldsymbol{X}) = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \to N(0;1)$

3. Předpoklady testu nelze popřít.

4. $W^* = \{x_{OBS}\colon\ x_{OBS} > z_{0,99}\} = \{x_{OBS}\colon\ x_{OBS} > 2{,}33\}$

5. $x_{OBS} = T(\boldsymbol{X})|H_0 = \frac{102-100}{12}\sqrt{100} = 1{,}67$

6. $x_{OBS} \notin W^* \Rightarrow$ Na hladině významnosti 0,01 nezamítáme H_0. Pozorované zvýšení průměrné výšky asijských hybridů lilií není na hladině významnosti 0,01 statisticky významné.

p-hodnota

- Nevýhodou klasického testu je skutečnost, že při pohledu na výsledek testu (vztah pozorované a kritické hodnoty) nevidíme přímo, jak rozhodnutí závisí na změně hladiny významnosti.

V appletu p-hodnota pozorujte vliv volby hladiny významnosti na rozhodnutí o výsledku testu při dané pozorované hodnotě.

Závěr:

V současnosti preferujeme rozhodování o výsledku testu na základě p-hodnoty, přičemž

p-hodnota je nejvyšší hladina významnosti, na níž se již nulová hypotéza nezamítá.

Čistý test významnosti aneb testování pomocí p-hodnoty

Postup při čistém testu významnosti

1. Formulace nulové a alternativní hypotézy.
2. Volba testové statistiky (testového kritéria) $T(\boldsymbol{X})$.
3. Ověření předpokladů testu.
4. Výpočet pozorované hodnoty x_{OBS} testové statistiky $T(\boldsymbol{X})$.
5. Výpočet p - hodnoty (angl. „p - value“ nebo „significance level“).

Tvar alternativní hypotézy H_A	$p\text{−}hodnota$
$\theta < \theta_0$	$p\text{−}hodnota = F_0(x_{OBS})$
$\theta > \theta_0$	$p\text{−}hodnota = 1 - F_0(x_{OBS})$
$\theta \neq \theta_0$	$p\text{−}hodnota = 2min\{F_0(x_{OBS}); 1 - F_0(x_{OBS})\}$

6. Rozhodnutí o výsledku testu.

$p\text{−}hodnota$	Rozhodnutí
$p\text{−}hodnota < \alpha$	Zamítáme H_0 ve prospěch H_A.
$p\text{−}hodnota \geq \alpha$	Nezamítáme H_0.

Výšku asijských hybridů lilií lze modelovat **náhodnou veličinou s normálním rozdělením** $N(100; 144)$. Skupina 100 kusů těchto lilií byla pěstována za příznivějších podmínek, aby se zjistilo, zda se výška zvýší.

b) Ověřte čistým testem významnosti, zda lze se spolehlivostí 0,95, resp. 0,99, tvrdit, že nové pěstební podmínky vedly ke zvýšení střední výšky asijských hybridů lilií.

Řešení:

„P-value is low,
null hypothesis must go!“
Keith M. Bower

1. H_0: $\mu = 100$, $\quad H_A$: $\mu > 100$

2. $T(\boldsymbol{X}) = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \to N(0;1)$

3. Předpoklady testu nelze popřít.

4. $x_{OBS} = T(\boldsymbol{x})|H_0 = \frac{102{,}5-100}{12}\sqrt{100} = 1{,}67$

5. $p - hodnota = 1 - F_0(1{,}67) = 1 - \Phi(1{,}67) = 0{,}047$

6. $p - hodnota < 0{,}05 \;\Rightarrow$ Na hladině významnosti 0,05 zamítáme H_0 ve prospěch H_A. Pozorované zvýšení průměrné výšky asijských hybridů lilií je na hladině významnosti 0,05 statisticky významné.

Výšku asijských hybridů lilií lze modelovat **náhodnou veličinou s normálním rozdělením** $N(100; 144)$. Skupina 100 kusů těchto lilií byla pěstována za příznivějších podmínek, aby se zjistilo, zda se výška zvýší.

b) Ověřte čistým testem významnosti, zda lze se spolehlivostí 0,95, resp. 0,99, tvrdit, že nové pěstební podmínky vedly ke zvýšení střední výšky asijských hybridů lilií.

Řešení:

„P-value is low,
null hypothesis must go!“
Keith M. Bower

1. $H_0\colon \mu = 100, \quad H_A\colon \mu > 100$

2. $T(\boldsymbol{X}) = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \to N(0; 1)$

3. Předpoklady testu nelze popřít.

4. $x_{OBS} = T(\boldsymbol{x})|H_0 = \frac{102{,}5-100}{12}\sqrt{100} = 1{,}67$

5. $p - hodnota = 1 - F_0(1{,}67) = 1 - \Phi(1{,}67) = 0{,}047$

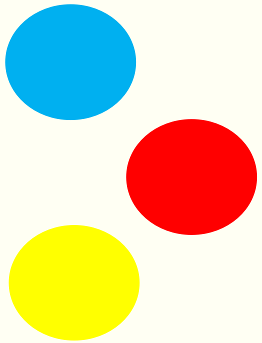
6. $p - hodnota > 0{,}01 \;\Rightarrow$ Na hladině významnosti 0,01 nezamítáme H_0. Pozorované zvýšení průměrné výšky asijských hybridů lilií není na hladině významnosti 0,01 statisticky významné.

Několik poznámek pro praxi

- Pozor na pečlivé plánování experimentu! (Nutno zajistit nezávislost pokusů, eliminaci vlivů nežádoucích faktorů, dostatečný rozsah výběru (výsledky testu nelze upravovat tím, že dodatečně rozšíříme výběrový soubor), ...)

Příklad: Včely jsou postupně vypouštěny do pokusného prostoru se žlutými, červenými a modrými terči. Sledujeme barvu terče, na který každá včela poprvé usedne. Nulová hypotéza je, že pravděpodobnost usednutí nezávisí na barvě terče (tímto způsobem zjišujeme, zda se včely vizuálně orientují a zda při této orientaci hrají nějakou úlohu barvy).

(Lepš, Kapitola 2 – testování hypotéz, test dobré shody, online: http://botanika.bf.jcu.cz/suspa/vyuka/materialy/KAP2.pdf [2012-03-19])



Co všechno je třeba při pokusu zajistit?

- vypouštění včel po jednotlivcích,
- včely nesmí zanechávat stopy o své návštěvě terče (není-li splněno, nutná výměna terčů po každém pokusu),
- předem daný počet pokusů.

Několik poznámek pro praxi

- V odborných pracích většinou výsledky testování hypotéz prezentujeme ve tvaru: (název testu, pozorovaná hodnota testové statistiky, p-hodnota) + někdy doplňujeme další parametry testu (např. stupně volnosti u χ^2 testu dobré shody). Např.: Předpoklad normality byl na hladině významnosti 0,05 zamítnut (χ^2 test dobré shody; χ^2=14,9; $df = 6; p - hodnota = 0{,}021$).

- Pokud je výstupem software: $p - hodnota = 0$, znamená to, že $p - hodnota$ je menší než přesnost software. Do odborných textů nikdy NEPIŠTE $p - hodnota = 0$, ale např. $p - hodnota << 0{,}01$.

- Uvědomte si, co to znamená, když $p - hodnota \to 1$! (Příliš dobrá shoda dat s nulovou hypotézou je podezřelá – angl. „Too good to be true!“)

DĚKUJI ZA POZORNOST!