Jan Vybíral

Lékařské testy individuální a skupinové

Rozhledy matematicko-fyzikální, Vol. 94 (2019), No. 1, 10–22

Persistent URL: `http://dml.cz/dmlcz/147678`

Chcete-li si své řešení nechat zkontrolovat, pošlete je na emailovou adresu autorky: lubomira.dvorakova@fjfi.cvut.cz

Literatura

[1] Joyce, J.: *Ulysses.* Sylvia Beach's Shakespeare and Company, Paris, 1922.

[2] Ryzák:, D.: *Antipalindromy.* SOČ práce, obor Matematika a statistika, roč. 40 (2018),
dostupné z: http://www.soc.cz/archiv-minulych-rocniku/

[3] on-line encyclopedia of integer sequences: https://oeis.org/A002385

Lékařské testy individuální a skupinové

Jan Vybíral, FJFI ČVUT, Praha

Abstrakt. Lékařské testy zná z běžného života každý z nás. Díky jejich nedokonalosti – vyjádřené pojmy senzitivita a specificita – se jejich vyhodnocení neobejde bez pojmů z elementární statistiky. Metoda skupinových testů (*group testing*), vyvinutá během druhé světové války, je oproti tomu stále aktivní pole vědeckého zájmu; původní článek Roberta Dorfmana z roku 1943 zaznamenává stále přes čtyřicet citací ročně. V této práci se pokusíme seznámit čtenáře se základy tohoto oboru, a to včetně teoretických i praktických cvičení.

1. Klasické lékařské testy

Diagnostika nemocí je prováděna lékařskými testy. Reálné testy ale nejsou ideální – ideální test by každého nemocného označil za nemocného a každého zdravého za zdravého. U každého v praxi používaného testu je tedy možné, že zdravý jedinec bude bohužel označen za nemocného (tzv. falešná pozitivita) nebo že nemocný jedinec bude označen za zdravého (tzv. falešná negativita).

Senzitivitou testu (neboli citlivostí testu) se nazývá pravděpodobnost, že nemocný pacient bude označen jako nemocný. Specificita testu je naopak pravděpodobnost, že zdravý jedinec bude označen za zdravého. Ideální test by tedy měl mít senzitivitu i specificitu rovnu jedné.

Cvičení 1.

1. (Hypotetický) test na rakovinu plic má senzitivitu 95 % a specificitu 98 %. Interpretujte tyto dva údaje.
2. Přibližně 1,5 % mužů ve věku 50 let trpí rakovinou plic. Jaká je pravděpodobnost, že padesátiletý muž s pozitivním testem na rakovinu plic jí opravdu trpí?
3. V populaci žen ve věku 50 let je zastoupení nemocných s rakovinou plic nižší, asi 0,5 %. Jaká je v tomto případě pravděpodobnost, že padesátiletá žena s pozitivním testem opravdu trpí rakovinou plic?
4. Je známo několik tisíc nemocí, které se vyskytují jen zřídka (tzv. vzácná onemocnění) a které postihují méně než 0,05 % populace (tedy nejvýše jednu ze dvou tisíc osob). Předpokládejme, že nějaká vzácná nemoc postihuje „jen" 0,02 % populace a že testy na tuto nemoc mají stejné parametry (tedy senzitivitu i specificitu) jako výše. Jaká je nyní pravděpodobnost, že pozitivně testovaná osoba touto nemocí opravdu trpí?

Simpsonův paradox[1)] upozorňuje na jeden jednoduchý efekt, který může snadno vést ke špatné interpretaci dat z klinických testů (a ostatně i jakýchkoliv jiných dat). Řekněme, že ledvinové kameny je možné léčit dvěma způsoby – postup A zahrnuje všechny druhy operací, tedy složité invazivní zákroky. Postup B spočívá v mnohem šetrnější terapii punkcí. Při jednom klinickém testu ([1]) byly sledovány různé případy pacientů s ledvinovými kameny (rozdělené do dvou skupin podle velikosti – malé a velké) a úspěšnost jednotlivých metod. Léčba A byla úspěšná v 81 z 87 případů léčby malých kamenů a v 192 z 263 v léčbě velkých kamenů. U léčby B byla tato úspěšnost 234 z 270 u malých kamenů a 55 z 80 u velkých kamenů. Protože

$$93\ \% \doteq \frac{81}{87} > \frac{234}{270} \doteq 87\ \% \quad \text{a} \quad 73\ \% \doteq \frac{192}{263} > \frac{55}{80} \doteq 69\ \%,$$

zdá se být léčba A úspěšnější jak pro léčbu malých, tak léčbu velkých kamenů. Protože ale

$$78\ \% = \frac{81+192}{87+263} < \frac{234+55}{270+80} \doteq 83\ \%,$$

1) Nejedná se o Homera Simpsona, ale o Edwarda Hugha Simpsona.

zdá se být léčba B paradoxně úspěšnější v léčbě ledvinových kamenů bez ohledu na velikost.

Rozdíl je samozřejmě dán tím, že méně invazivní metoda byla často volena pro léčbu malých kamenů, kde jsou ale obě metody podstatně úspěšnější než při léčbě velkých kamenů. Ať jde o malé nebo velké kameny, je podle těchto dat úspěšnější metoda A. Přesvědčení, že celkově je lepší metoda B, není správné. Sčítat počty úspěchů u různě závažných případů ($81 + 192$, resp. $234 + 55$) je zavádějící.

2. Skupinové testy

V průběhu druhé světové války se americká armáda potýkala s doslova delikátním problémem (popsaným v [2]). Bylo zapotřebí podrobit lékařským testům velký počet vojáků.[2)] Teoreticky sice stačilo každému vojáku odebrat vzorek krve a otestovat každý vzorek zvlášť, ale tyto testy byly finančně nákladné a v daném množství nerealizovatelné. Navíc se dalo předpokládat, že nemocných vojáků bude jen mizivé procento ze všech testovaných. Tak vznikla metoda skupinového testování, tzv. *group testing.*

Uvažujme skupinu N jedinců (tedy oněch vojáků), které je třeba otestovat. Někteří z nich jsou zdraví, někteří nemocní. Pokud si je nějakým způsobem seřadíme (např. podle abecedy), tak jejich zdravotní stav je možné popsat posloupností Z, která má N nul a jedniček. Tedy např. pokud $N = 5$ a $Z = (0, 1, 0, 0, 1)$, tak druhý a pátý voják jsou nakažení, ostatní jsou zdraví. Posloupnost Z na začátku testování neznáme a testováním ji chceme přesně určit. Pro jednoduchost budeme v této části předpokládat, že prováděné testy jsou ideální, tedy že senzitivita i specificita testů je rovna jedné.

Metoda skupinového testování je založena na nápadu, kdy vybereme nějakou podskupinu vojáků, smícháme jejich vzorky (nebo alespoň části jejich vzorků) a otestujeme tuto směs. Pokud jsou testy dostatečně senzitivní, tak bude možné tímto způsobem zjistit, zda alespoň jeden z vojáků dané podskupiny byl nakažen. Tím tedy můžeme "otestovat" hned několik vojáků najednou; při negativním testu budeme vědět, že jsou všichni zdraví, ale při pozitivním testu se pouze dozvíme, že některý z nich byl nakažený. Nebude tedy možné s jistotou říct, který z nich byl nakažen. To bude možné přesně říct až po dalších testech.

[2)] V naší úloze to nebude hrát žádnou roli, ale není důvod zapírat, že hledanou nemocí byl syfilis.

Cvičení 2.

1. Předpokládejme, že předem víme, že z dané skupiny N vojáků je nejvýše jeden nakažený. Jak máme volit výběr podskupin, abychom s co nejmenším počtem testů zjistili, který (pokud vůbec nějaký) je nakažený?
2. Je vámi nalezený postup optimální (tj. je počet testů skutečně minimální)?

Tento postup lze sice zobecnit, i pokud víme, že počet nakažených vojáků je nejvýše k, kde $k \geq 1$, navržená metoda má ale jednu zásadní nevýhodu. Tou je neúměrná velikost vybraných podskupin – až $N/2$. To klade nerealistické nároky na senzitivitu testů. Jedno z možných řešení tohoto problému je dáno poněkud překvapivým přístupem, kdy členy testované podskupiny vybereme náhodně.

Zvolme tedy nejprve reálný parametr $0 < p < 1$ a vyberme do testované skupiny každého vojáka s pravděpodobností p. Matematicky můžeme tento postup modelovat pomocí nezávislých náhodných proměnných x_i, $1 \leq i \leq N$,

$$x_i = \begin{cases} 1 & \text{s pravděpodobností} \quad p, \\ 0 & \text{s pravděpodobností} \quad 1 - p. \end{cases}$$

Pokud $x_i = 1$, tak i-tý voják bude do výběru zařazen; pokud $x_i = 0$, do výběru zařazen nebude.

Cvičení 3.

1. Jaký je průměrný počet vojáků zařazených do testovacího vzorku?
2. Jaká je pravděpodobnost, že vybraných vojáků bude více (nebo méně) než n, kde n je nějaké přirozené číslo mezi 1 a N?
3. Jaký je rozptyl počtu vybraných vojáků?
4. Pro $N = 10^5$ a $p = 0{,}01$ udělejte histogram počtu náhodně vybraných vojáků při $R = 1$, $R = 10^3$, $R = 10^5$ a $R = 10^7$ nezávislých opakováních výběru. Porovnejte tyto výsledky vzájemně a s vypočtenou pravděpodobností.

Celé testování velkého počtu vojáků samozřejmě nebude možné provést jedním výběrem a jedním testem, jak jsme ostatně viděli už ve cvičení 2. Proces výběru a testování tedy bude nutné několikrát opakovat – počet těchto opakování označme přirozeným číslem m. Celý test

N vojáků proběhne tedy odběrem vzorků všech vojáků a výběrem m podskupin, jejichž vzorky budou smíchány a otestovány. Nakonec bude naším úkolem navrhnout algoritmus, který zkombinuje informaci o výběrech a jejich výsledcích a rozhodne, které vojáky test označí za nakažené a které za zdravé.

Cvičení 4.

1. Navrhněte matematické označení pro výběr m testovacích skupin.
2. Do kolika testovacích vzorků bude každý voják v průměru zařazen?
3. Pokud nějaký voják nebude nikdy vybrán, nebude z výsledků testů nikdy možné určit, zda byl, nebo nebyl nakažen. Jaká je pravděpodobnost, že jeden pevně zvolený voják nebude nikdy vybrán?
4. Jaká je pravděpodobnost, že alespoň jeden z vojáků nebude zařazen do žádného výběru?

Poté, co provedeme m výběrů a testů, musíme rozhodnout, který z vojáků je či není nakažený. Navržený algoritmus je velice jednoduchý: Voják bude celým testovacím procesem označen za nakaženého, pokud všechny testy, do nichž byl vybrán, vyjdou pozitivně.[3)]

Cvičení 5.

1. Jaká je pravděpodobnost, že navržený testovací postup s parametry k, m, N, p funguje korektně, tedy že pro každého vojáka bylo správně určeno, zda je, či není nakažený?
2. Pro dané $\varepsilon > 0$ určete, kolik opakování testu bude stačit provést, má-li být pravděpodobnost úspěchu celého algoritmu větší než $1 - \varepsilon$.
3. Výsledek ověřte numerickou simulací – pro $N = 100\,000$, $k = 20$, $p = 0{,}02$ a různá m zjistěte empirickou pravděpodobnost, jak často bude s těmito parametry daný test úspěšný.

3. Návody a řešení

V této sekci naznačíme možné postupy řešení úloh a cvičení.

[3)] V souladu s obvyklou konvencí, kdy prázdný obecný kvantifikátor je pravdivý, bude voják, který nebyl zařazen do žádného výběru, označen také za nakaženého.

Cvičení 1:

1. Test odhalí nemocného pacienta s pravděpodobností 95 % a zdravého jedince neoznačí jako nemocného s pravděpodobností 98 %.
2. Úlohu lze řešit (alespoň) dvěma, navzájem ovšem podobnými způsoby.

   a) Nemocných mužů daného věku je 1,5 % a 95 % z nich má pozitivní test. Celkem trpí tedy $0{,}95 \cdot 0{,}015 = 0{,}01425$, tj. 1,425 % mužů ve věku 50 let rakovinou plic a má i pozitivní test. Naopak, zdravých mužů je 98,5 % a 2 % z nich mají pozitivní test. Tedy $0{,}985 \cdot 0{,}02 = 0{,}0197$, tj. 1,97 % mužů daného věku má pozitivní test, ale rakovinou plic netrpí. Celkem má tedy 3,395 % všech padesátiletých mužů pozitivní test. Ze všech mužů s pozitivním testem jich ale rakovinou plic trpí pouze $1{,}425/3{,}395 \doteq 0{,}42 = 42$ %. Hledaná pravděpodobnost je tedy 42 %.

   b) Úlohu lze řešit také použitím pojmu podmíněné pravděpodobnosti. Máme-li dva jevy A a B, pak pravděpodobnost, že nastane A, pokud nastává B, označíme

   $$\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}.$$

   Tuto definici je možné použít, jen pokud $\mathsf{P}(B) > 0$. Podmíněná pravděpodobnost tedy udává, v kolika procentech případů nastává jev A, uvažujeme-li celkově jen případy, kdy už víme, že nastává jev B.

   Označme X zdravotní stav padesátiletého muže. Tedy $X = 1$, pokud má daný muž rakovinu plic, a $X = 0$, pokud je zdráv. V případě, že testy provádíme na náhodně vybraných mužích z celé populace[4)], můžeme ztotožnit X s náhodnou proměnnou, pro kterou platí $\mathsf{P}(X = 1) = \pi = 0{,}015$ a $\mathsf{P}(X = 0) = 1 - \pi = 0{,}985$. Dále označíme T výsledek testu padesátiletého muže, tedy $T = 1$, pokud je testovaný muž testem označen za nemocného (tj. test je pozitivní), a $T = 0$, pokud test je negativní. I veličinu T tedy můžeme interpretovat jako náhodnou proměnnou, která ale není

[4)] Tento předpoklad není téměř nikdy splněn dokonale – obvykle jsou testováni lidé podle nějakého klíče či předvýběru; například proto, že u nich existuje podezření na danou nemoc, nebo proto, že se přihlásili do klinické studie.

nezávislá na X. Parametry testu lze pomocí tohoto označení přepsat jako

$$\mathsf{P}(T=1|X=1)=\frac{\mathsf{P}(T=1\ \&\ X=1)}{\mathsf{P}(X=1)}=p=0{,}95,$$

$$\mathsf{P}(T=0|X=0)=\frac{\mathsf{P}(T=0\ \&\ X=0)}{\mathsf{P}(X=0)}=q=0{,}98.$$

Úlohou je poté spočítat $\mathsf{P}(X=1|T=1)$. Nejprve ovšem vypočteme (srovnejte s předcházejícím postupem!)

$$\mathsf{P}(T=1\ \&\ X=1)=\frac{\mathsf{P}(T=1\ \&\ X=1)}{\mathsf{P}(X=1)}\cdot\mathsf{P}(X=1)=p\pi$$

a z $\mathsf{P}(X=0)=\mathsf{P}(X=0\ \&\ T=0)+\mathsf{P}(X=0\ \&\ T=1)$ obdržíme

$$\begin{aligned}\mathsf{P}(T=1\ \&\ X=0)&=\mathsf{P}(X=0)-\mathsf{P}(T=0\ \&\ X=0)=\\&=1-\pi-\frac{\mathsf{P}(T=0\ \&\ X=0)}{\mathsf{P}(X=0)}\mathsf{P}(X=0)=\\&=1-\pi-(1-\pi)q=(1-\pi)(1-q).\end{aligned}$$

Tyto vztahy použijeme k finálnímu výpočtu:

$$\begin{aligned}\mathsf{P}(X=1|T=1)&=\frac{\mathsf{P}(X=1\ \&\ T=1)}{\mathsf{P}(T=1)}=\frac{p\pi}{\mathsf{P}(T=1)}=\\&=\frac{p\pi}{\mathsf{P}(T=1\ \&\ X=1)+\mathsf{P}(T=1\ \&\ X=0)}=\\&=\frac{p\pi}{p\pi+(1-q)(1-\pi)}.\end{aligned}$$

Dosazení $\pi=0{,}015$, $p=0{,}95$ a $q=0{,}98$ dává opět 42 %.

3. Dosazení $\pi=0{,}005$, $p=0{,}95$ a $q=0{,}98$ dává 19,3 %. Padesátiletá žena s pozitivním testem trpí tedy rakovinou plic s pravděpodobností 19,3 %.

4. Pro $\pi=0{,}0002$, $p=0{,}95$ a $q=0{,}98$ dostaneme 0,94 %, tedy méně než jedno procento.

Na závěr ještě jedno varování k interpretaci výsledků tohoto cvičení. Navržené výpočty jsou správné za předpokladu, že testy byly prováděny osobám náhodně vybraným z celé populace, tedy při plošném screeningu. Rakovinové (i jiné) testy jsou ale obvykle prováděny především těm, u kterých existuje nějaké podezření. Celkově vzato bude tedy pravděpodobnost, že daná osoba při pozitivním testu danou nemocí opravdu trpí, asi vyšší než uvádějí předchozí cvičení.

Cvičení 2: Metodou půlení intervalu lze najít potenciálního nakaženého při $\lceil \log_2(N) \rceil$ testech. Pokud ale provedeme testů méně, tak bude buď existovat voják, který nebyl zařazen ani do jedné testovací skupiny, nebo budou existovat dva vojáci, kteří budou při každém testu testováni společně. Pokud je tedy nakažený jen jeden z nich, menší počet testů nebude stačit.

Cvičení 3:

1. Střední hodnotu počtu vybraných vojáků lze spočítat různými způsoby. Každá z náhodných proměnných $x_1, \dots, x_N$ má střední hodnotu

   $$1 \cdot p + 0 \cdot (1 - p) = p.$$

   Počet vybraných vojáků je dán součtem $x_1 + \dots + x_N$ a jeho průměrná hodnota je součtem průměrných hodnot, tedy pN.

   Druhý způsob využívá binomické věty a definice střední hodnoty – a je podstatně pracnější. Nejprve si uvědomíme, že pravděpodobnost, že vybraných vojáků bude právě l je

   $$\binom{N}{l} p^l (1-p)^{N-l} \quad \text{pro } l = 0, \dots, N.$$

   Mimochodem, z binomické věty pak plyne, že

   $$\sum_{l=0}^{N} \binom{N}{l} p^l (1-p)^{N-l} = (p + (1-p))^N = 1,$$

   což odpovídá tomu, že vybraný počet vojáků musí být právě jedno celé číslo od nuly do N. Střední hodnota počtu vybraných vojáků je potom (podle definice) rovna

$$\sum_{l=0}^{N} l \cdot \mathsf{P}(\text{bylo vybráno právě } l \text{ vojáků}) =$$

$$= \sum_{l=0}^{N} \binom{N}{l} p^l (1-p)^{N-l} \cdot l = \sum_{l=1}^{N} \frac{N!}{(l-1)!(N-l)!} p^l (1-p)^{N-l} =$$

$$= pN \sum_{l=0}^{N-1} \binom{N-1}{l} p^l (1-p)^{N-l-1} = pN(p + (1-p))^{N-1} = pN.$$

2. Pravděpodobnost, že vybraných vojáků bude alespoň n je

$$\sum_{l=n}^{N} \binom{N}{l} p^l (1-p)^{N-l}.$$

3. Úlohu lze opět řešit alespoň dvěma způsoby. Rozptyl x_i je (z definice) roven

$$p(1-p)^2 + (1-p)(-p)^2 = p(1-p).$$

   Rozptyl součtu nezávislých(!) náhodných proměnných je roven součtu jejich rozptylů,[5] výsledek je tedy $Np(1-p)$.

   Nebo (a opět podstatně pracněji) se lze dopracovat téhož opět užitím definice a binomické věty, kterýžto postup přenecháme pilnému čtenáři.

Cvičení 4:

1. Pro $1 \leq i \leq N$ a $1 \leq j \leq m$ budeme uvažovat nezávislé proměnné x_i^j, kde opět $x_i^j = 1$ s pravděpodobností p a $x_i^j = 0$ s pravděpodobností $1-p$. (Zdůrazněme, že j je zde horní index a ne mocnina.) Do j-tého výběru budou zařazeni vojáci s pořadovými čísly $\{i\colon x_i^j = 1\}$.
2. Pravděpodobnost zařazení daného vojáka do jednoho výběru je p, celkem tedy bude v průměru zařazen do pm výběrů.
3. Pravděpodobnost, že daný voják nebude zařazen do žádného výběru je $(1-p)^m$.

[5] Předpoklad nezávislosti je nutný, bez něj toto tvrzení neplatí.

4. Spočítat danou pravděpodobnost je sice možné z principu inkluze a exkluze, ale velice technické a výsledná formule je prakticky těžko použitelná. Danou pravděpodobnost lze ale snadno omezit shora – tedy najít poměrně jednoduchý výraz, jehož hodnota bude vždy o něco málo vyšší. Označme jako X_i událost „i-tý voják nebyl zařazen do žádného výběru“, tedy $x_i^1 = \ldots = x_i^m = 0$. Událost X – aspoň jeden voják nebyl zařazen do žádného výběru – je pak sjednocením $X = X_1 \cup \ldots \cup X_N$, a tedy

$$\mathsf{P}(X) \leq \sum_{i=1}^{N} \mathsf{P}(X_i) = N(1-p)^m.$$

Pokud je výraz napravo větší než jedna, je tento odhad triviální – každá pravděpodobnost je nejvýše rovna jedné. Pro parametry, které budou později důležité, je to však vcelku dobrý odhad. Chceme-li například, aby $\mathsf{P}(X) \leq \varepsilon < 1$, stačí volit m tak velké, aby

$$N(1-p)^m < \varepsilon,$$

tedy

$$m > \frac{\ln(\varepsilon) - \ln(N)}{\ln(1-p)} = \frac{\ln(N) + \ln(1/\varepsilon)}{\ln(1/(1-p))}.$$

Při pevném ε a p roste tedy počet nutných opakování výběru m pouze logaritmicky s N – metoda se tedy zdá být opravdu vhodná pro velké počty testovaných jedinců.

Cvičení 5:

1. Úlohu vyřešíme postupně v několika krocích.

   (a) Pokud je nějaký voják nakažený, tak všechny testy, do kterých byl vybrán, vyšly pozitivně. Může se ale stát, že nějaký zdravý voják byl v testu vždy s nějakým nakaženým vojákem a všechny testy, do kterých byl vybrán, tak vyšly pozitivně. Může tedy dojít k falešné pozitivitě.

   (b) Označme $\tau(i)$ počet testovacích skupin, do kterých byl vybrán i-tý voják. Jedná se tedy o náhodnou veličinu, která nabývá hodnot od 0 do m. Má binomické rozdělení, tedy platí

$$\mathsf{P}(\tau(i) = l) = \binom{m}{l} p^l (1-p)^{m-l}, \quad i = 1, \ldots, N,\ l = 0, \ldots, m.$$

(c) Pravděpodobnost, že do jednoho výběru nebyl zařazen žádný nakažený voják je $(1-p)^{\varkappa} \geq (1-p)^k$, kde $\varkappa \leq k$ je počet nakažených vojáků. Naopak pravděpodobnost, že do výběru byl zařazen aspoň jeden nakažený voják, je $1-(1-p)^{\varkappa} \leq 1-(1-p)^k$. Vezměme jedno i pevné a nechť $\tau(i) = l$; tedy i-tý voják byl vybrán do l testů. Dále předpokládejme, že tento voják nakažený nebyl. Pak pravděpodobnost, že ve všech těchto l testech byl i nějaký nakažený voják, je nejvýše $[1-(1-p)^k]^l$.

(d) Připomeňme, že nás zajímají jen zdraví vojáci, protože pouze u nich může vést postup k chybě. Předpokládejme tedy, že i-tý voják je zdravý. Označme A_i tu událost, že s i-tým vojákem byl v testu vždy i nějaký jiný nakažený voják. Pokud nastane A_i, tak všechny testy, kterých se i-tý voják účastnil, vyšly pozitivně, a my ho tudíž falešně označíme za nemocného. Zbývá odhadnout pravděpodobnost této události:

$$\begin{aligned}\mathsf{P}(A_i) &= \sum_{l=0}^{m} \mathsf{P}(A_i \mathbin{\&} \tau(i) = l) = \sum_{l=0}^{m} \mathsf{P}(A_i | \tau(i) = l)\mathsf{P}(\tau(i) = l) \leq \\ &\leq \sum_{l=0}^{m} \left[1-(1-p)^k\right]^l \binom{m}{l} p^l (1-p)^{m-l} = \\ &= \sum_{l=0}^{m} \binom{m}{l} \left[p(1-(1-p)^k)\right]^l (1-p)^{m-l} = \\ &= \left(p - p(1-p)^k + 1 - p\right)^m = \left(1 - p(1-p)^k\right)^m.\end{aligned}$$

(e) Tento odhad platí pro každého zdravého vojáka, kterých je nejvýše N. Pravděpodobnost, že tedy existuje aspoň jeden voják, se kterým byl v každém testu i nějaký nakažený voják, je tedy opět nejvýše N-krát větší, tedy

$$N\left(1 - p(1-p)^k\right)^m.$$

2. Má-li být daný výraz menší než $\varepsilon > 0$, potřebujeme zvolit parametry testu tak, aby
$$N\left(1 - p(1-p)^k\right)^m < \varepsilon$$
tj.
$$\ln(N) + m \ln\left[1 - p(1-p)^k\right] < \ln(\varepsilon),$$

neboli
$$m > \frac{\ln(N) + \ln(1/\varepsilon)}{-\ln(1 - p(1-p)^k)}.$$

3. Pro dané parametry a $\varepsilon = 0{,}01$ vychází cca $m = 1\,200$, tedy přibližně osmdesátkrát méně testů, než kdybychom testovali každého vojáka individuálně. Velikost testovacích skupin je v průměru $pN = 2\,000$.

   V programu Matlab lze jeden test se zvolenými parametry N, m, p a k simulovat například následujícím skriptem:

```
=======================
N=100000; % pocet vojaku
m=1200; % pocet mereni
p=0.02; % pravdepodobnost zahrnuti do vzorku
k=20; % maximalni pocet nakazenych vojaku

Z=zeros(N,1); % vektor oznacujici zdravotni stav vojaku
for j=1:k
   Z(randi(N,1))=1;
end

X=binornd(1,p,[m,N]);
% mxN matice, kazdy radek je jedno mereni

mereni=sign(X*Z); % vysledky jednotlivych mereni

Y=sign(transpose(X)*(mereni-ones(m,1)))+1;
% vojaci oznaceni testem za nakazene

if Y == Z % pokud Z=Y, test byl uspesny, jinak ne
   test=1;
else
   test=0;
end
=======================
```

   Spuštěním skriptu ve smyčce je dále možné experimentálně ověřit četnost úspěšnosti či neúspěšnosti testu. Při pěti průchodech s $R = 1000$ opakováními bylo po řadě dosaženo 986, 990, 995, 994 a 991 úspěšných testů, což je ve výborné shodě s předchozím výpočtem.

Literatura

[1] Charig, C. R., Webb, D. R., Payne, S. R., Wickham, J. E.: Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British Medical Journal*, roč. 292 (1986), s. 879–882.

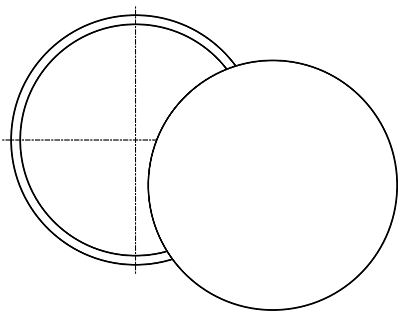
[2] Dorfman, R.: The detection of defective members of large populations. *The Annals of Mathematical Statistics*, roč. 14 (1943), s. 436–440.

Proč jsou kanály kulaté

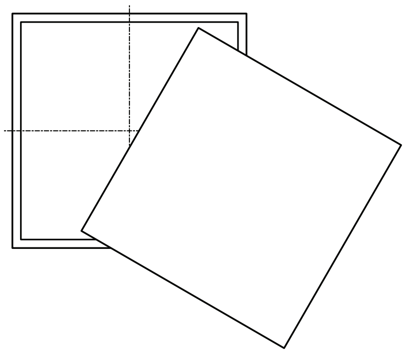
Eduard Šubert, Praha

Viděli jste někdy na ulici víko od kanálu? A bylo kulaté? Ale proč by mělo být kulaté?

Kruh má totiž takovou praktickou vlastnost: konstantní průměr. Nehledě na to, jak kruhové víko otočíte, do kruhového kanálu nemůže nikdy spadnout (obr. 1a).



(a) Kruh



(b) Čtverec

Obr. 1: Jaké víko může spadnout do kanálu?

Naproti tomu třeba takové čtvercové víko má stranu kratší než úhlopříčku. Když tedy víko nevhodně otočíte, může se propadnout (obr. 1b). Nejen, že byste pak museli víko z kanálu vytáhnout, ale mohlo by spadnout na hlavu dělníkovi, který se vydal opravovat kanalizaci, a to je