Zkušenost: Statistická indukce

CORE004 Matematika jako součást kultury

Zdeněk Pospíšil
707@mail.muni.cz

Masarykova univerzita

11. listopadu 2021

Obsah

Princip maximální věrohodnosti

„Jest zcela nezpochybnitelným faktem, že nemůžeme-li poznat nejpravdivější soudy, musíme se řídit soudy nejpravděpodobnějšími."

René Descartes, Rozprava o metodě

Princip maximální věrohodnosti

Aplikace: Odhad počtu jedinců volně žijící populace

Lincolnova-Petersonova metoda, mark-recatch

Princip maximální věrohodnosti

Aplikace: Odhad počtu jedinců volně žijící populace

Lincolnova-Petersonova metoda, mark-recatch

- Odchytíme a označkujeme m jedinců.
- Po dostatečném čase, ale ne příliš dlouhém, odchytíme dostatečné množství dalších jedinců a spočítáme počet označených mezi nimi.

Princip maximální věrohodnosti

Aplikace: Odhad počtu jedinců volně žijící populace

počet nově ulovených jedinců: r počet označených jedinců mezi nimi: k
neznámý počet jedinců v populaci: n

Princip maximální věrohodnosti

Aplikace: Odhad počtu jedinců volně žijící populace

počet nově ulovených jedinců: r počet označených jedinců mezi nimi: k
neznámý počet jedinců v populaci: n; určitě je $n \geq \max\{m, r\}$

Princip maximální věrohodnosti

Aplikace: Odhad počtu jedinců volně žijící populace

počet nově ulovených jedinců: r počet označených jedinců mezi nimi: k
neznámý počet jedinců v populaci: n; určitě je $n \geq \max\{m, r\}$

počet možností, jak mezi n jedinci vybrat r: $c(n, r) = \binom{n}{r}$

Princip maximální věrohodnosti

Aplikace: Odhad počtu jedinců volně žijící populace

počet nově ulovených jedinců: r počet označených jedinců mezi nimi: k
neznámý počet jedinců v populaci: n; určitě je $n \geq \max\{m, r\}$

počet možností, jak mezi n jedinci vybrat r: $c(n, r) = \binom{n}{r}$

počet možností, jak z m označených jedinců vybrat k: $c(m, k) = \binom{m}{k}$

Princip maximální věrohodnosti

Aplikace: Odhad počtu jedinců volně žijící populace

počet nově ulovených jedinců: r počet označených jedinců mezi nimi: k
neznámý počet jedinců v populaci: n; určitě je $n \geq \max\{m, r\}$

počet možností, jak mezi n jedinci vybrat r: $c(n,r) = \binom{n}{r}$

počet možností, jak z m označených jedinců vybrat k: $c(m,k) = \binom{m}{k}$

počet možností, jak z $n-m$ neoznačených jedinců vybrat $r-k$:

$$c(n-m, r-k) = \binom{n-m}{r-k}$$

Princip maximální věrohodnosti

Aplikace: Odhad počtu jedinců volně žijící populace

počet nově ulovených jedinců: r počet označených jedinců mezi nimi: k
neznámý počet jedinců v populaci: n; určitě je $n \geq \max\{m, r\}$

počet možností, jak mezi n jedinci vybrat r: $c(n,r) = \binom{n}{r}$

počet možností, jak z m označených jedinců vybrat k: $c(m,k) = \binom{m}{k}$

počet možností, jak z $n-m$ neoznačených jedinců vybrat $r-k$:

$$c(n-m, r-k) = \binom{n-m}{r-k}$$

A_n^{mrk} ... jev: populace je tvořena n jedinci, mezi nimi je m označených a při druhém odchytu mezi r ulovenými jedinci bylo k označených

Princip maximální věrohodnosti

Aplikace: Odhad počtu jedinců volně žijící populace

počet nově ulovených jedinců: r počet označených jedinců mezi nimi: k
neznámý počet jedinců v populaci: n; určitě je $n \geq \max\{m, r\}$

počet možností, jak mezi n jedinci vybrat r: $c(n, r) = \binom{n}{r}$

počet možností, jak z m označených jedinců vybrat k: $c(m, k) = \binom{m}{k}$

počet možností, jak z $n - m$ neoznačených jedinců vybrat $r - k$:

$$c(n - m, r - k) = \binom{n-m}{r-k}$$

A_n^{mrk} ... jev: populace je tvořena n jedinci, mezi nimi je m označených a při druhém odchytu mezi r ulovenými jedinci bylo k označených

$$\mathrm{P}(A_n^{mrk}) = \frac{\binom{m}{k}\binom{n-m}{r-k}}{\binom{n}{r}}$$

Princip maximální věrohodnosti

Aplikace: Odhad počtu jedinců volně žijící populace

počet nově ulovených jedinců: r počet označených jedinců mezi nimi: k
neznámý počet jedinců v populaci: n; určitě je $n \geq \max\{m, r\}$

počet možností, jak mezi n jedinci vybrat r: $c(n,r) = \binom{n}{r}$

počet možností, jak z m označených jedinců vybrat k: $c(m,k) = \binom{m}{k}$

počet možností, jak z $n-m$ neoznačených jedinců vybrat $r-k$:

$$c(n-m, r-k) = \binom{n-m}{r-k}$$

A_n^{mrk} … jev: populace je tvořena n jedinci, mezi nimi je m označených a při druhém odchytu mezi r ulovenými jedinci bylo k označených

$$\mathrm{P}(A_n^{mrk}) = \frac{\binom{m}{k}\binom{n-m}{r-k}}{\binom{n}{r}}$$

Hledáme takové n, aby při daných m, r, k byla hodnota $\mathrm{P}(A_n^{mrk})$ maximální.

Princip maximální věrohodnosti

Aplikace: Odhad počtu jedinců volně žijící populace

počet nově ulovených jedinců: r počet označených jedinců mezi nimi: k
neznámý počet jedinců v populaci: n; určitě je $n \geq \max\{m, r\}$

počet možností, jak mezi n jedinci vybrat r: $c(n, r) = \binom{n}{r}$

počet možností, jak z m označených jedinců vybrat k: $c(m, k) = \binom{m}{k}$

počet možností, jak z $n - m$ neoznačených jedinců vybrat $r - k$:

$$c(n - m, r - k) = \binom{n - m}{r - k}$$

A_n^{mrk} ... jev: populace je tvořena n jedinci, mezi nimi je m označených a při druhém odchytu mezi r ulovenými jedinci bylo k označených

$$\mathrm{P}(A_n^{mrk}) = \frac{\binom{m}{k}\binom{n-m}{r-k}}{\binom{n}{r}}$$

Hledáme takové n, aby při daných m, r, k byla hodnota $\mathrm{P}(A_n^{mrk})$ maximální.

Je $n \in \left\langle \frac{mr}{k}, \frac{mr}{k} + 1 \right)$.

Rozložení náhodné veličiny

Konkrétní případ: binomické rozdělení (Bernoulliovo)

Pravděpodobnost úspěchu v nějakém pokusu je rovna p.
Pokus zopakujeme n-krát. Jaká je pravděpodobnost jevu B_n^k, že úspěch nastane právě k-krát?

Počet možností výběru k pořadových čísel úspěšných pokusů mezi n provedenými:

$$c(n,k) = \binom{n}{k}$$

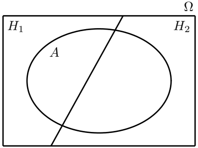
Pravděpodobnost k úspěchů a $n-k$ neúspěchů:

$$p^k(1-p)^{n-k}$$

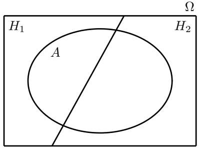
Celkem:

$$\mathrm{P}(B_n^k) = \binom{n}{k} p^k(1-p)^{n-k}$$

Bayesův vzorec



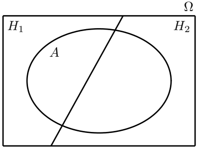
Bayesův vzorec



Inverzní pravděpodobnost:

$$\mathrm{P}(H|A) = \frac{\mathrm{P}(H \cap A)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A \cap H)\mathrm{P}(H)}{\mathrm{P}(H)\mathrm{P}(A)} = \frac{\mathrm{P}(H)}{\mathrm{P}(A)}\mathrm{P}(A|H)$$

Bayesův vzorec



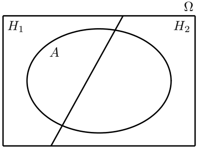
Inverzní pravděpodobnost:

$$\mathrm{P}(H|A) = \frac{\mathrm{P}(H \cap A)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A \cap H)\mathrm{P}(H)}{\mathrm{P}(H)\mathrm{P}(A)} = \frac{\mathrm{P}(H)}{\mathrm{P}(A)}\mathrm{P}(A|H)$$

Celková pravděpodobnost:

$$\begin{aligned}\mathrm{P}(A) &= \mathrm{P}\big((A \cap H_1) \cup (A \cap H_2)\big) = \mathrm{P}(A \cap H_1) + \mathrm{P}(A \cap H_2) = \\ &= \mathrm{P}(H_1)\mathrm{P}(A|H_1) + \mathrm{P}(H_2)\mathrm{P}(A|H_2)\end{aligned}$$

Bayesův vzorec



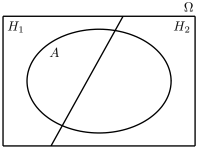
Inverzní pravděpodobnost:

$$\mathrm{P}(H|A) = \frac{\mathrm{P}(H)}{\mathrm{P}(A)}\mathrm{P}(A|H)$$

Celková pravděpodobnost:

$$\mathrm{P}(A) = \mathrm{P}(H_1)\mathrm{P}(A|H_1) + \mathrm{P}(H_2)\mathrm{P}(A|H_2)$$

Bayesův vzorec



Inverzní pravděpodobnost:

$$\mathsf{P}(H|A) = \frac{\mathsf{P}(H)}{\mathsf{P}(A)}\mathsf{P}(A|H)$$

Celková pravděpodobnost:

$$\mathsf{P}(A) = \mathsf{P}(H_1)\mathsf{P}(A|H_1) + \mathsf{P}(H_2)\mathsf{P}(A|H_2)$$

$$\mathsf{P}(H_1|A) = \frac{\mathsf{P}(H_1)\mathsf{P}(A|H_1)}{\mathsf{P}(H_1)\mathsf{P}(A|H_1) + \mathsf{P}(H_2)\mathsf{P}(A|H_2)}$$

Bayesův vzorec

Aplikace: Diagnostické testy

Vyšetřovaný objekt buď zjišťovanou charakteristiku (chorobu) má nebo nemá.
Test je postup, který dá právě jeden ze dvou možných výsledků:

$+$ pozitivní, choroba zjištěna
$-$ negativní, choroba nezjištěna

Bayesův vzorec

Aplikace: Diagnostické testy

Vyšetřovaný objekt buď zjišťovanou charakteristiku (chorobu) má nebo nemá.
Test je postup, který dá právě jeden ze dvou možných výsledků:

$+$ pozitivní, choroba zjištěna
$-$ negativní, choroba nezjištěna

Vlastnosti testu:

Senzitivita – podíl pozitivních výsledků testu mezi všemi výsledky při testování osob, které sledovanou charakteristiku (chorobu) mají.

Specificita – podíl negativních výsledků testu mezi všemi výsledky při testování osob, které sledovanou charakteristiku (chorobu) nemají.

Bayesův vzorec

Aplikace: Diagnostické testy

Vyšetřovaný objekt buď zjišťovanou charakteristiku (chorobu) má nebo nemá.
Test je postup, který dá právě jeden ze dvou možných výsledků:

$+$ pozitivní, choroba zjištěna
$-$ negativní, choroba nezjištěna

Vlastnosti testu:

Senzitivita – podíl pozitivních výsledků testu mezi všemi výsledky při testování osob, které sledovanou charakteristiku (chorobu) mají.

Specificita – podíl negativních výsledků testu mezi všemi výsledky při testování osob, které sledovanou charakteristiku (chorobu) nemají.

Předpokládáme, že známe pravděpodobnost výskytu zjišťované charakteristiky.

incidence … počet nových případů choroby za časovou jednotku
prevalence … celkový počet nemocných

Bayesův vzorec

Aplikace: Diagnostické testy

Vyšetřovaný objekt buď zjišťovanou charakteristiku (chorobu) má nebo nemá.
Test je postup, který dá právě jeden ze dvou možných výsledků:

$+$ pozitivní, choroba zjištěna
$-$ negativní, choroba nezjištěna

Vlastnosti testu:

Senzitivita – podíl pozitivních výsledků testu mezi všemi výsledky při testování osob, které sledovanou charakteristiku (chorobu) mají.

Specificita – podíl negativních výsledků testu mezi všemi výsledky při testování osob, které sledovanou charakteristiku (chorobu) nemají.

Předpokládáme, že známe pravděpodobnost výskytu zjišťované charakteristiky.

incidence … počet nových případů choroby za časovou jednotku
prevalence … celkový počet nemocných

Otázka: jaká je pravděpodobnost, že vyšetřovaný objekt příslušnou charakteristiku vykazuje (má chorobu), pokud test dal pozitivní výsledek? $\mathrm{P}(H|+) =?$

Bayesův vzorec

Aplikace: Diagnostické testy

Vyšetřovaný objekt buď zjišťovanou charakteristiku (chorobu) má nebo nemá.
Test je postup, který dá právě jeden ze dvou možných výsledků:

$+$ pozitivní, choroba zjištěna
$-$ negativní, choroba nezjištěna

Vlastnosti testu:

Senzitivita – podíl pozitivních výsledků testu mezi všemi výsledky při testování osob, které sledovanou charakteristiku (chorobu) mají. $p = \mathrm{P}(+|H)$

Specificita – podíl negativních výsledků testu mezi všemi výsledky při testování osob, které sledovanou charakteristiku (chorobu) nemají. $q = \mathrm{P}(-|H')$

Předpokládáme, že známe pravděpodobnost výskytu zjišťované charakteristiky $r = \mathrm{P}(H)$.
incidence … počet nových případů choroby za časovou jednotku
prevalence … celkový počet nemocných

Otázka: jaká je pravděpodobnost, že vyšetřovaný objekt příslušnou charakteristiku vykazuje (má chorobu), pokud test dal pozitivní výsledek? $\mathrm{P}(H|+) = ?$

Označení jevů: H … objekt vykazuje charakteristiku (má chorobu)
H' … objekt nevykazuje charakteristiku (nemá chorobu)

Bayesův vzorec

Aplikace: Diagnostické testy

Vyšetřovaný objekt buď zjišťovanou charakteristiku (chorobu) má nebo nemá.
Test je postup, který dá právě jeden ze dvou možných výsledků:

$+$ pozitivní, choroba zjištěna
$-$ negativní, choroba nezjištěna

Vlastnosti testu:

Senzitivita – podíl pozitivních výsledků testu mezi všemi výsledky při testování osob, které sledovanou charakteristiku (chorobu) mají. $p = \mathrm{P}(+|H)$

Specificita – podíl negativních výsledků testu mezi všemi výsledky při testování osob, které sledovanou charakteristiku (chorobu) nemají. $q = \mathrm{P}(-|H')$

Předpokládáme, že známe pravděpodobnost výskytu zjišťované charakteristiky $r = \mathrm{P}(H)$.
incidence … počet nových případů choroby za časovou jednotku
prevalence … celkový počet nemocných

Otázka: jaká je pravděpodobnost, že vyšetřovaný objekt příslušnou charakteristiku vykazuje (má chorobu), pokud test dal pozitivní výsledek? $\mathrm{P}(H|+) = ?$

Označení jevů: H … objekt vykazuje charakteristiku (má chorobu)
H' … objekt nevykazuje charakteristiku (nemá chorobu)

Jevy $+$ a $-$, H a H' jsou komplementární:

$$\mathrm{P}(H') = 1 - r,\ \mathrm{P}(+|H') = 1 - q,\ \mathrm{P}(-|H) = 1 - p$$

Bayesův vzorec

Aplikace: Diagnostické testy

$$
\begin{array}{lll}
\mathrm{P}(H) = r & \mathrm{P}(+|H) = p & \mathrm{P}(-|H) = 1 - p \\
\mathrm{P}(H') = 1 - r & \mathrm{P}(+|H') = 1 - q & \mathrm{P}(-|H') = q
\end{array}
$$

Bayesův vzorec

Aplikace: Diagnostické testy

$$
\begin{array}{lll}
\mathrm{P}(H) = r & \mathrm{P}(+|H) = p & \mathrm{P}(-|H) = 1 - p \\
\mathrm{P}(H') = 1 - r & \mathrm{P}(+|H') = 1 - q & \mathrm{P}(-|H') = q
\end{array}
$$

Bayesův vzorec:

$$
\mathrm{P}(H|+) = \frac{\mathrm{P}(H)\mathrm{P}(+|H)}{\mathrm{P}(H)\mathrm{P}(+|H) + \mathrm{P}(H')\mathrm{P}(+|H')} = \frac{rp}{rp + (1-q)(1-r)} = \frac{rp}{1 - r - q + rp + rq}
$$

$$
\mathrm{P}(H|-) = \frac{\mathrm{P}(H)\mathrm{P}(-|H)}{\mathrm{P}(H)\mathrm{P}(-|H) + \mathrm{P}(H')\mathrm{P}(-|H')} = \frac{r(1-p)}{r(1-p) + (1-r)q} = \frac{r(1-p)}{r + q - rq - rp}
$$

Bayesův vzorec

Aplikace: Diagnostické testy

$$\begin{array}{lll} \mathrm{P}(H) = r & \mathrm{P}(+|H) = p & \mathrm{P}(-|H) = 1-p \\ \mathrm{P}(H') = 1-r & \mathrm{P}(+|H') = 1-q & \mathrm{P}(-|H') = q \end{array}$$

Bayesův vzorec:

$$\mathrm{P}(H|+) = \frac{rp}{1-r-q+rp+rq}$$

$$\mathrm{P}(H|-) = \frac{r(1-p)}{r+q-rq-rp}$$

Bayesův vzorec

Aplikace: Diagnostické testy

$$
\begin{array}{lll}
\mathrm{P}(H) = r & \mathrm{P}(+|H) = p & \mathrm{P}(-|H) = 1-p \\
\mathrm{P}(H') = 1-r & \mathrm{P}(+|H') = 1-q & \mathrm{P}(-|H') = q
\end{array}
$$

Bayesův vzorec:

$$\mathrm{P}(H|+) = \frac{rp}{1-r-q+rp+rq}$$

$$\mathrm{P}(H|-) = \frac{r(1-p)}{r+q-rq-rp}$$

Příklad: AIDS	incidence	$r = 0,001$
	senzitivita	$p = 0{,}998$
	specificita	$q = 0{,}99$

Bayesův vzorec

Aplikace: Diagnostické testy

$$\begin{array}{lll} \mathrm{P}(H) = r & \mathrm{P}(+|H) = p & \mathrm{P}(-|H) = 1 - p \\ \mathrm{P}(H') = 1 - r & \mathrm{P}(+|H') = 1 - q & \mathrm{P}(-|H') = q \end{array}$$

Bayesův vzorec:

$$\mathrm{P}(H|+) = \frac{rp}{1 - r - q + rp + rq}$$

$$\mathrm{P}(H|-) = \frac{r(1-p)}{r + q - rq - rp}$$

Příklad: AIDS	incidence	$r = 0,001$	
	senzitivita	$p = 0{,}998$	
	specificita	$q = 0{,}99$	$\mathrm{P}(H\|+) = 0{,}091$

Bayesův vzorec

Aplikace: Diagnostické testy

$$
\begin{array}{lll}
\mathrm{P}(H) = r & \mathrm{P}(+|H) = p & \mathrm{P}(-|H) = 1-p \\
\mathrm{P}(H') = 1-r & \mathrm{P}(+|H') = 1-q & \mathrm{P}(-|H') = q
\end{array}
$$

Bayesův vzorec:

$$\mathrm{P}(H|+) = \frac{rp}{1-r-q+rp+rq}$$

$$\mathrm{P}(H|-) = \frac{r(1-p)}{r+q-rq-rp}$$

Příklad: AIDS	incidence	$r = 0,001$	
	senzitivita	$p = 0{,}998$	
	specificita	$q = 0{,}99$	$\mathrm{P}(H\|+) = 0{,}091$

Kumulace zkušenosti: osoby s pozitivním výsledkem otestujeme znovu

Bayesův vzorec

Aplikace: Diagnostické testy

$$\begin{array}{lll} P(H) = r & P(+|H) = p & P(-|H) = 1 - p \\ P(H') = 1 - r & P(+|H') = 1 - q & P(-|H') = q \end{array}$$

Bayesův vzorec:

$$P(H|+) = \frac{rp}{1 - r - q + rp + rq}$$

$$P(H|-) = \frac{r(1-p)}{r + q - rq - rp}$$

Příklad: AIDS	incidence	$r = 0,001$	
	senzitivita	$p = 0{,}998$	
	specificita	$q = 0{,}99$	$P(H|+) = 0{,}091$

Kumulace zkušenosti: osoby s pozitivním výsledkem otestujeme znovu

$$r_1 = 0{,}091 : \qquad P(H|++) = 0{,}909$$

Bayesův vzorec

Aplikace: Diagnostické testy

$$\begin{array}{lll} P(H) = r & P(+|H) = p & P(-|H) = 1-p \\ P(H') = 1-r & P(+|H') = 1-q & P(-|H') = q \end{array}$$

Bayesův vzorec:

$$P(H|+) = \frac{rp}{1-r-q+rp+rq}$$

$$P(H|-) = \frac{r(1-p)}{r+q-rq-rp}$$

Příklad: AIDS	incidence	$r = 0,001$	
	senzitivita	$p = 0{,}998$	
	specificita	$q = 0{,}99$	$P(H|+) = 0{,}091$

Kumulace zkušenosti: osoby s pozitivním výsledkem otestujeme znovu

$r_1 = 0{,}091$: $\quad P(H|++) = 0{,}909$

$r_2 = 0{,}909$: $\quad P(H|+++) = 0{,}999, \; P(H|++-) = 0{,}020$

MASARYKOVA UNIVERZITA