## Primeri uporabe verjetnosti v medicini in biologiji

Ogledali si bomo nekaj preprostih zgledov iz medicine in biologije, kjer lahko uporabimo osnovno znanje o (diskretni) verjetnosti. Kasneje bomo na take zglede še naleteli.

### Krvni testi

Pri ugotavljanju prisotnosti virusne ali bakterijske okužbe uporabljajo različne krvne teste, ki testirajo kri glede prisotnosti protitelesc. Test je dober, če z veliko verjetnostjo pokaže na prisotno bolezen. Naj bo $B$ dogodek, da je bolezen (okužba) prisotna, in $A$ dogodek, da je test pozitiven (pokaže na prisotnost bolezni). Zanima nas torej pogojna verjetnost $P(A|B)$. Običajno je to zelo visoka vrednost, npr. 0.95, kar pomeni, da je test ob prisotnosti okužbe pozitiven s 95% verjetnostjo. To število imenujemo *občutljivost testa*.

Pri vsakem testiranju lahko pride do napake. V načelu sta možni dve vrsti napak:
1. *Napaka prve vrste* nastopi, če je test negativen, čeprav je bolezen prisotna. Verjetnost za to napako je (po formuli za verjetnost nasprotnega dogodka) $P(A^c|B) = 1 - P(A|B) \approx 1 - 0.95 = 0.05$.
2. *Napaka druge vrste* nastopi, če je test pozitiven, čeprav bolezni ni. Verjetnost za to napako $P(A|B^c)$ je težje oceniti, med zanesljivo znano populacijo bi morali oceniti verjetnost pozitivnega testa. Običajno je tudi to zelo majhna vrednost, npr. spet približno 0.05.

Problem, ki se realno vedno postavlja pred zdravnika, ki izvaja test, pa je ocena pogojne verjetnosti $P(B|A)$, da ima pacient res bolezen, če je bil test pozitiven. Test namreč ni 100%. To verjetnost ocenimo po Bayesovi obratni formuli:

$$P(B|A) = \frac{P(B)P(A|B)}{P(B)P(A|B) + P(B^c)P(A|B^c)}.$$

Očitno moramo, poleg pogojnih verjetnosti na desni strani te formule, ki smo jih že ocenili, poznati tudi brezpogojno verjetnost $P(B)$, da je bolezen prisotna (verjetnost nasprotnega dogodka $P(B^c)$, da bolezni ni, potem takoj izračunamo). Ocena za $P(B)$ je zelo zahtevna, videli pa bomo, da je od nje močno odvisna željena ocena $P(B|A)$.

Pri redkih boleznih (v normalni populaciji), kot je npr. okužba z virusom HIV, ki povzroča AIDS, je verjetnost bolezni zelo majhna, statistično npr. pod 0.006 (po podatkih WHO iz leta 1988). Drugo je npr. testiranje rizičnih skupin ali populacije v nekaterih afriških predelih. Za naše namene vzemimo npr., da je $P(B) \approx 0.003$. Potem dobimo po Bayesu

$$P(B|A) \approx \frac{0.003 \cdot 0.95}{0.003 \cdot 0.95 + 0.997 \cdot 0.05} \approx 0.054.$$

Ta rezultat je po svoje presenetljiv: v skoraj 95% primerov bolezen ni prisotna, čeprav je test bil pozitiven.

Kaj pa, če bi bolezen ne bila tako redka, ampak bi npr. vzeli $P(B) \approx 0.1$? Po isti formuli bi sedaj izračunali $P(B|A) \approx 0.68$. Pogojna verjetnost, da je bolezen prisotna, če je bil test pozitiven, je zdaj močno narasla, s prejšnjih 5% na več kot 2/3.

Matematična razlaga je preprosta. Pogojna verjetnost $P(A|B)$ je racionalna funkcija spremenljivke $x = P(B)$ na desni strani Bayesove formule, torej (pri nespremenjenih ostalih podatkih v našem primeru)

$$P(B|A) = \frac{0.95x}{0.95x + 0.5(1-x)} = \frac{0.95x}{0.90x + 0.05} = \frac{95x}{90x + 5} = \frac{19x}{18x + 1}.$$

Če bi jo narisali, bi videli, da je njen graf v okolici točke 0 zelo strm (odvod v 0 je enak 19), zato hitro narašča oziroma je zelo občutljiva že na malenkostno povečanje vrednosti za $x$.

Vse to samo pomeni, da se pri redkih boleznih ne splača izvajati množičnih (preventivnih) testov, saj ni zanesljiv (kljub pozitivnosti je veliko več možnosti, da bolezni ni). Prihranjeni čas in denar je bolje uporabiti za testiranje rizičnih skupin, kjer je pomoč bolj potrebna.

Oglejmo si še eno, včasih precej pogosto nepravilno uporabo obratne formule pri ugotavljanju očetovstva.

ZGLED. Neki moški je bil osumljen očetovstva. Pri moškem so ugotovili zelo redko gensko lastnost (genetski znak), za katerega se ve, da se s 100% zanesljivostjo prenaša z očeta na otroka. Otroka so testirali in tudi pri njem našli genetski znak. Kolikšna je verjetnost, da je moški res oče danega otroka?

Za oceno te verjetnosti bi bilo smiselno spet uporabiti Bayesovo obratno formulo. Naj bo $B$ dogodek, da je moški dejansko oče otroka in $A$ dogodek, da pri otroku najdemo genetski znak. Potem je $P(A|B) = 1$, saj se znak zanesljivo prenese z očeta na otroka, in $P(A|B^c) \approx 0.01$, saj je tolikšna verjetnost spontanega posedovanja genetskega znaka v celotni populaciji. Kot običajno je problem oceniti vrednost $P(B)$, od katere je pogojna verjetnost, kot smo videli prej, močno odvisna. Toda zdaj tega ne moremo določiti statistično. Moški je namreč oče ali pa ni, poskusa ne moremo ponavljati v nedogled. Pred isto dilemo je sodnik, ki naj o primeru odloči. Zanesti se mora na svojo subjektivno oceno in na njeni podlagi določiti verjetnost. Če vzamemo vrednost $P(B) \approx 0.5$ (enake šanse, da je ali da ni oče), potem dobimo po obratni formuli:

$$P(B|A) \approx \frac{0.5 \cdot 1}{0.5 \cdot 1 + 0.5 \cdot 0.01} = \frac{500}{505} = \frac{100}{101} \approx 0.99,$$

torej kar 99% verjetnost, da je moški oče. Izhajali smo iz ad hoc sprejete ocene o 50% verjetnosti očetovstva in naenkrat dobili 99% verjetnost očetovstva. To gotovo ni prav, saj smo iz dejansko nevednosti (zato smo se odločili za 50% možnost) dobili visoko in obtožujočo vrednost 99%. Upoštevajmo še znano dejstvo, da je pred zakonom vsak nedolžen, dokler mu ni krivda dokazana, pa zlahka sprevidimo absurdnost dobljenega rezultata. Z obratno formulo ni nič narobe, napačna je (v tem primeru) njena uporaba.

**Štetje živali**

V splošnem izračunamo verjetnost, da se v $n$ ponovitvah poskusa dogodek z verjetnostjo $p$ zgodi natanko $k$-krat, po *Bernoullijevi formuli*:

$$P_n(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Tu je lahko $k = 0, 1, 2, ..., n$.

ZGLED. V rezervatu živi $a$ levinj in $b$ leva. Verjetnost, da pri obhodu vidimo posamezno levinjo je torej $p = a/(a+b)$, leva pa $1 - p = b/(a+b)$. Verjetnost, da pri $n$ opažanjih posamezne mačke vidimo $k$ levinj (lahko istih) je po Bernoullijevi formuli $P_n(k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} a^k b^{n-k}/(a+b)^n$. Če je npr. $a = 3$ in $b = 2$, je $p = 3/5$ in $1 - p = 2/5$. Pri petih opažanjih je verjetnost, da vidimo 3 levinje, enaka $P_5(3) = 10 \cdot 3^3 \cdot 2^2/5^5 = 34,5\%$.

V zgornjem zgledu se je popolnoma isti dogodek lahko večkrat ponovil, npr. večkrat smo lahko opazili *isto* levinjo. To je tako, kot če bi iz posode z belimi in črnimi kroglicami večkrat zapored na slepo potegnili kroglico in jo potem spet vedno vrnili v posodo. Pri tem beležimo, kolikokrat v $n$ poskusih, potegnemo belo kroglico. Rečemo, da gre za *izbiranje z vračanjem*. Kadar pa imamo *izbiranje brez vračanja*, kroglic vmes ne vračamo. Kolikšna je verjetnost, da v $n$ ponovitvah $k$-krat potegnemo belo, če je v posodi z $N$ kroglicami $K$ belih in $N - K$ črnih? To moramo izračunati drugače, tako kot da bi $n$ kroglic potegnili

naenkrat in med njimi našli $k$ belih. Zdaj velja formula

$$P_n(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

ZGLED. Denimo, da v rezervatu zagledamo skupino 3 velikih mačk. Kolikšna je verjetnost, da sta med njimi 2 levinji (in 1 lev)? Zdaj imamo $P_3(2) = \binom{3}{2}\binom{2}{1}/\binom{5}{3} = 3\cdot 2/10 = 3/5$.

Da se pokazati, da pri je velikem $N, K$ in $N-K$ zadnja formula (pri izbiranju brez vračanja) približno enaka prejšnji (pri izbiranju z vračanjem), kjer pa vzamemo $p = K/N$.

Formulo za izračun verjetnosti pri izbiranju brez vračanja izkoristimo za oceno števila rib v velikem ribniku.

ZGLED. V ribniku je neznano število rib. Njihovo število $N$ ocenimo na naslednji način. Najprej ulovimo $K$ rib, jih označimo in spustimo nazaj v ribnik. Čez nekaj časa, ko se ribe dobro premešajo, ponovno ujamemo nekaj rib, recimo $n$, in med njimi najdemo $k$ označenih. Izračunamo verjetnost, da se to zgodi po formuli za $P_n^N(k)$ (z vračanjem), in poiščemo tak $N$, da je verjetnost $P_n^N(k)$ največja. To je hkrati najbolj verjetno število rib v ribniku.

Kako pa poiščemo $\max P_n^N(k)$? Tako, da med seboj primerjamo $P_n^N(k)$ in $P_n^{N+1}(k)$. Po kratkem računu dobimo

$$P_n^{N+1}(k)/P_n^N(k) = \frac{\binom{N+1-K}{n-k}\binom{N}{n}}{\binom{N+1}{n}\binom{N-K}{n-k}}$$

$$= \frac{(N+1-K)(N+1-n)}{(N+1-K-n+k)(N+1)},$$

kar je $> 1$ natanko takrat, ko je $nK > (N+1)k$. Najverjetnejše število je torej $N \sim nK/k$. To je ocena za število rib v ribniku.

**Hardy-Weinbergov zakon v genetiki**

Z osnovami verjetnosti, ki smo jih spoznali doslej, lahko preprosto ocenimo razmerje med različnimi genotipi skozi zaporedne generacije. Navedimo dva primera, ki mo ju že obravnavali.

1. Predpostavimo, da opazujemo en sam lokus z dvema aleloma $A$ in $a$. Denimo, da imamo v začetni generaciji $u$ genotipov $AA$, $2v$ genotipov $Aa$ in $w$ genotipov $aa$. Ker je število vseh oseb $u + 2v + w$, so torej verjetnosti posameznih genotipov $P(AA) = u/(u+2v+w)$, $P(Aa) = 2v/(u+2v+w)$ in $P(aa) = w/(u+2v+w)$. V procesu spolne delitve (meiosis), se pari alelov razdružijo in posamezni aleli iščejo nove partnerske alele za genetsko rekombinacijo. Izračunajmo verjetnosti posameznih alelov. Po formuli za polno verjetnost je $P(A) = P(AA)\cdot 1 + P(Aa)\cdot 1/2 = (u+v)/(u+2v+w)$ in $P(a) = 1 - P(A) = (v+w)/(u+2v+w)$. Označimo $p = P(A)$ in $q = P(a)$ (seveda je $p+q = 1$).

Ker se prosti aleli kombinirajo neodvisno, izračunamo verjetnosti posameznih genotipov v naslednji generaciji po Bernoulijevi formuli (na dveh mestih se na vsakem zgodi $A$ ali $a$). Torej: $P(AA) = p^2$, $P(Aa) = 2pq$ in $P(aa) = q^2$. V naslednji meiozi, dobimo (spet po formuli za polno verjetnost) $p_1 = p^2 + pq = p$ in $q_1 = pq + q^2 = q$ (isto kot prej) in zato ponovno za verjetnosti genotipov v novi generaciji

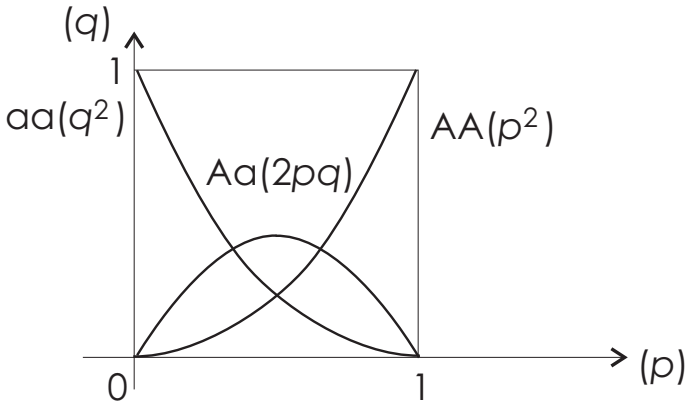
$$P(AA) = p^2, \quad P(Aa) = 2pq \ \text{ in } \ P(aa) = q^2.$$

Iz generacije v generacijo isto. Torej se verjetnosti (se pravi deleži) posameznih genotipov iz generacijo v generacijo ohranjajo. To je t.i. *Hardy-Weinbergov* genetski zakon, imenovan po angleškem matematiku Hardyju in zdravniku Weinbergu.

Zakon lahko nazorno predstavimo s t.i. *Punnettovim kvadratom*

| | | | mati | |
|---|---|---|---|---|
| | | $A$ | | $B$ |
| | | $p$ | | $q$ |
| oče | $A$ | $p$ | $p^2$ | $pq$ |
| | $B$ | $q$ | $pq$ | $q^2$ |

ali z grafom, ki prikazuje odvisnost verjetnosti za tri možne genotipe od spremenljivke $p$ $(0 < p < 1)$, ki pomeni verjetnost gena $A$.




Slika 1

2. Podobno situacijo bi imeli pri genih z več aleli na določenem lokusu, npr. pri krvnih skupinah: če so verjetnosti za krvno skupino $A$, $B$ in 0 po vrsti $p$, $q$ in $r$, so verjetnosti posameznih genotipov naslednje: $P(AA) = p^2$, $P(BB) = q^2$, $P(00) = r^2$, $P(AB) = 2pq$, $P(A0) = 2pr$ in $P(B0) = 2qr$. To se ponavlja iz roda v rod, tako da ostanejo deleži genotipov in zato tudi deleži krvnih skupin iz generacijo v generacijo konstantni. Za Anglijo so v zvezi s krvnimi grupami znana naslednja razmerja (glej [**?**], str. 122): A 32.1%, B 22.4%, AB 7.1%, 0 38.4%. Preverimo, ali je to skladno z modelom Hardyja in Weinberga. V tem modelu so verjetnosti za posamezne krvne skupine naslednje: $p_A = P(AA) + P(A0) = p^2 + 2pr$, $p_B = P(BB) + P(B0) = q^2 + 2qr$, $p_{AB} = P(AB) = 2pq$ in $p_0 = P(00) = r^2$. Hitro izračunamo, da mora biti $p = \sqrt{p_A + p_0} - \sqrt{p_0} \approx 0.22$, $q = \sqrt{p_B + p_0} - \sqrt{p_0} \approx 0.16$ in $r = \sqrt{p_0} \approx 0.62$.

3. Včasih je kakšna različica gena (kakšen okvarjen gen) lahko usodna za človeka, čeprav je recesivna. Denimo, da je $A$ zdrava, $a$ pa okvarjena različica gena. Genotip $aa$ pomeni bolezen. Denimo, da je ta bolezen tako huda, da človek, ki jo podeduje, zboli in umre že v otroštvu, tj. predno odraste in ima lahko otroke. Kljub temu lahko v sebi nosi okvarjen gen, in sicer v primeru, ko ima genotip $Aa$ (medtem ko je $AA$ genotip popolnoma zdravega človeka). Denimo, da je v normalni populaciji verjetnost nosilca bolezni enaka $P(Aa) = p$. Običajno gre za redko bolezen, zato je $p$ majhno število, blizu 0. Toda če imamo o nekem odraslem človeku podatek, da je eden od bratov ali sester umrl za to boleznijo, potem ta človek ni več poljuben predstavnik celotne populacije, temveč ima, kot rečemo, *zgodovino*. Verjetnost, da je nosilec bolezni, ni več tako majhna. Zdaj gre za pogojno verjetnost. Oba njegova starša sta morala biti nosilca bolezni, torej genotipa $Aa$, sicer ne bi mogla imeti potomca genotipa $aa$, ki je umrl. Ker je človek, ki ga raziskujemo, preživel otroštvo in odrastel, ne more biti genotipa $aa$, lahko le $AA$ ali $Aa$. Torej je pogojna verjetnost za

vsakega od teh dveh možnih genotipov enaka

$$P(AA/AA \cup Aa) = \frac{P(AA)}{P(AA)+P(Aa)} = \frac{1/4}{1/4+1/2} = 1/3 \text{ in}$$

$$P(Aa/AA \cup Aa) = \frac{P(Aa)}{P(AA)+P(Aa)} = \frac{1/2}{1/4+1/2} = 2/3.$$

Vidimo, da je (pogojna) verjetnost, da je obravnavana oseba nosilec bolezni, zdaj 2/3, kar je precej več kot pri človeku brez zgodovine, ko je ta verjetnost enaka p (blizu 0).

Kako pa je z njegovimi otroki? Denimo, da se poroči z žensko, za katero ne obstajajo podatki o bolnih sorodnikih. Ona je lahko genotipa $AA$ z verjetnostjo $1-p$ ali genotipa $Aa$ z verjetnostjo $p$. Verjetnost različnih kombinacij je naslednja:
$P(AA \times AA) = \frac{1}{3} \cdot (1-p) = (1-p)/3$; otrok je z verjetnostjo 1 genotipa $AA$;
$P(AA \times Aa) = \frac{1}{3} \cdot p = p/3$; otrok je z verjetnostjo 1/2 genotipa $AA$ in z verjetnostjo 1/2 genotipa $Aa$;
$P(Aa \times AA) = \frac{2}{3} \cdot (1-p) = 2(1-p)/3$; otrok je z verjetnostjo 1/2 genotipa $AA$ in z verjetnostjo 1/2 genotipa $Aa$;
$P(Aa \times Aa) = \frac{2}{3} \cdot p = 2p/3$; otrok je z verjetnostjo 1/4 genotipa $AA$, z verjetnostjo 1/2 genotipa $Aa$ in z verjetnostjo 1/4 genotipa $aa$.
Torej so za tega otroka za posamezen genotip značilne naslednje verjetnosti, izračunane po formuli za polno verjetnost glede na zgornje štiri primere:
$P_1(AA) = (1-p)/3 \cdot 1 + p/3 \cdot 1/2 + 2(1-p)/3 \cdot 1/2 + 2p/3 \cdot 1/4 = 2/3 - p/3$,
$P_1(Aa) = p/3 \cdot 1/2 + 2(1-p)/3 \cdot 1/2 + 2p/3 \cdot 1/2 = 1/3 + p/6$,
$P_1(aa) = 2p/3 \cdot 1/4 = p/6$.
Verjetnost, da otrok umre je samo $p/6$, da odraste $1-p/6$ in da je nosilec bolezni, če odraste

$$P_1(Aa/AA \cup Aa) = \frac{1/3+p/6}{1-p/6} = \frac{2+p}{6-p}.$$

Če je $p$ zanemarljiv, je to približno 1/3, torej pol manj kot pri njegovem očetu.

**Opomba**. Ker na razvoj bolezni vplivajo običajno tudi drugi genetski faktorji, velja zgornji izračun le za prvo in grobo aproksimacijo.

4. Predpostavimo, da so v populaciji trije genotipi $AA$, $Aa$ in $aa$ v razmerju $p^2 : 2pq : q^2$, $p+q=1$, kot določa Hardy-Weinbergov zakon. Slučajno izbrana starša imata otroka genotipa $AA$. Vemo, da je verjetnost za to tudi enaka $p^2$. Potem se odločita še za drugega otroka. Ta pa ni več neodvisno izbran, ima *zgodovino*. Verjetnost, da je tudi drugi otrok genotipa $AA$, je zdaj pogojna in odvisna od tega, katera dva starševska genotipa sta se srečala. Izračunati moramo $P((AA)_2/(AA)_1) = P((AA)_1 \cap (AA)_2)/P((AA)_1)$. Seveda je po Hardyju in Weinbergu imenovalec enak $p^2$, verjetnost v števcu pa izračunamo po formuli za polno verjetnost glede na kombinacije staršev:

$$P((AA)_1 \cap (AA)_2) = P(AA \times AA) \cdot 1^2 + P(AA \times Aa) \cdot 1/2^2 + P(Aa \times Aa) \cdot 1/4^2 = p^4 + p^3 q + p^2 q^2/4.$$

Dobimo

$$P((AA)_2/(AA)_1) = P((AA)_1 \cap (AA)_2)/P((AA)_1) = p^2 + pq + q^2/4 = (1+p)^2/4.$$

Podobno bi dobili tudi druge verjetnosti, npr.

$$P((Aa)_2/(Aa)_1) = P((Aa)_1 \cap (Aa)_2)/P((Aa)_1) = (p^2 + 3pq + q^2)/2 = (1+pq)/2.$$

## Literatura

[1] N.F. Britton, *Essential Mathematical Biology*, Springer Undergraduate Mathematics Series, 2003.