19 Verjetnostni račun

Preštevanje – Poskusi in izidi – Verjetnosti izidov – Verjetnost sestavljenih izidov – Binomska porazdelitev – Vsota slučajnih izidov – Normalna porazdelitev – Povprečje in varianca – Večdimenzijske porazdelitve – Soodvisnost spremenljivk – Vzorčenje in statistika – Merjenje in merske napake – Intervalno ocenjevanje – Preizkušanje domnev – Regresijska analiza – Statistično zavajanje

19.1 Preštevanje

Izbiranja

Nekatere stvari v življenju lahko naredimo na več načinov. Dober primer je kosilo v restavraciji. Na jedilniku je zapisano: 2 predjedi, 3 glavne jedi in 2 poobedka. Izberemo lahko po eno jed iz vsake skupine. Koliko različnih kosil si lahko privoščimo? Očitno $N = 2 \cdot 3 \cdot 2$. Nasploh velja: če lahko najprej naredimo N_1 izbir; nato – neodvisno od tega, kaj smo izbrali – novih N_2 izbir; in tako naprej, je različnih izbirnih nizov $N = N_1 \cdot N_2 \ldots N_n$. Kaže, da sta izbiranje in preštevanje izbir pomembni opravili. Poskusimo torej raziskati kaj več o tem.

Permutacije

Imejmo niz petih različnih črk $(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e})$. Ta niz lahko premešamo; ena izmed premešav je, na primer, $(\mathrm{b}, \mathrm{a}, \mathrm{c}, \mathrm{e}, \mathrm{d})$. Rečemo, da je to *permutacija* osnovnega niza. Koliko pa je takih različnih permutacij? Na prvo mesto permutacije lahko postavimo eno izmed 5 črk. Ostanejo še štiri. Na drugo mesto postavimo eno izmed preostalih 4 črk. Tako nadaljujemo in dobimo $N = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5!$ različnih nizov črk. Na splošno lahko torej iz n-terice različnih elementov naredimo P_n njenih permutacij:

$$P_n = n! \,. \tag{19.1}$$

Če vseh n elementov ni različnih, ampak je med njimi r enakih, je različnih permutacij $r!$-krat manj: ${P_n}^r = n!/r!$.

Variacije

Iz niza petih črk $(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e})$ potegnimo poljubne tri črke. Trojke iz istih črk, a z različnim vrstnim redom, obravnavamo kot različne: $(\mathrm{a}, \mathrm{b}, \mathrm{c})$ je torej različna od $(\mathrm{b}, \mathrm{a}, \mathrm{c})$. Rečemo, da so to *variacije* dolžine 3 iz osnovnega niza. Koliko različnih variacij pa lahko naredimo? Na prvo mesto v trojki lahko postavimo eno izmed 5 črk. Preostanejo štiri. Na drugo mesto postavimo eno izmed preostalih 4 črk. Tako nadaljujemo in dobimo $N = 5 \cdot 4 \cdot 3 = 5!/(5-3)!$ različnih trojk. Na splošno iz n-terice različnih elementov lahko naredimo ${V_n}^r$ različnih variacij dolžine r:

$${V_n}^r = \frac{n!}{(n-r)!} \,. \tag{19.2}$$

Kombinacije

Koliko je pa različnih trojk, pri čemer obravnavamo trojke iz istih črk, a z različnim vrstnim redom, kot enake: $(\mathrm{a}, \mathrm{b}, \mathrm{c})$ je enaka

(b, a, c)? Rečemo, da so to *kombinacije* dolžine 3 iz osnovnega niza. Očitno je število kombinacij manjše kot število variacij in sicer za tolikokrat, kolikor je permutacij niza z dolžino 3, torej $N = 5!/(5-3)!3!$. Na splošno lahko torej iz n-terice različnih elementov naredimo ${C_n}^r$ različnih kombinacij dolžine r:

$$ {C_n}^r = \frac{n!}{r!\,(n-r)!}\,. \tag{19.3} $$

19.2 Poskusi in izidi

Igralna kocka

Ljudje, ki nimajo kaj boljšega početi, radi mečejo kocke. Takšna *igralna kocka* ima na svojih ploskvah narisane pike. Vsaka ploskev ima svoje število pik: od ena do šest. Ko kocko vržemo na mizo, se zakotali, ustavi in njena zgornja ploskev pokaže določeno število pik. Vnaprej nikoli ne vemo, koliko jih bo padlo. Ljudje stavijo denar, kaj se bo pri metu zgodilo, in tisti, ki ugane, pobere stave. Te so lahko raznovrstne: padla bo trojka; ne bo padla trojka; padlo bo sodo število; v dveh zaporednih metih bo padla vsaj ena šestica; pri hkratnem metu dveh kock bo padlo skupaj deset pik; in še mnogo drugega.

Slika 19.1 Igralni kocki. Izid meta ene ali več kock je slučajna spremenljivka. (Anon)

Poskus in izid

Na met kocke lahko pogledamo kot na *poskus*, ki ima šest možnih *elementarnih izidov*: število pik od ena do šest. Vnaprej ne vemo, kakšen bo izid predstoječega poskusa, zato rečemo, da je tak izid *slučajna spremenljivka*, ki lahko zavzame celoštevilčne vrednosti med ena in šest. Pričakujemo pa, da se bo v velikem številu poskusov (torej metov), pojavil vsak izmed šestih izidov v približno enakem deležu in sicer v eni šestini primerov, če je le kocka "poštena". Pravzaprav je res obratno: če se vsak izid pojavlja enako pogosto, rečemo, da je kocka poštena.

19.3 Verjetnosti izidov

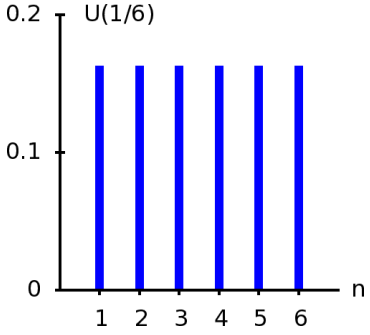
Pogostost izida

Pa izmerimo, kako pogosto se pojavljajo posamični izidi za dotično kocko! Kar naprej jo mečimo in beležimo vsakokratne izide, to je vrednosti slučajne spremenljivke x. Ta spremenljivka lahko zavzame vrednosti $x_1 = 1$, $x_2 = 2$ … $x_6 = 6$. Ko vržemo kocko 10-krat, se izid x_3, na primer, pojavi 2-krat, torej v 2/10 poskusov. Pri N poskusih se nasploh izid x_k pojavi N_k-krat. Razmerje N_k/N se z vsakim nadaljnjim metom spremeni. V začetku se od meta do meta močno spreminja, kasneje pa se čedalje bolj zgošča okrog

neke limitne vrednosti. Vsak izid se zgošča okrog svoje limite. S tem je definirana njegova *relativna frekvenca* oziroma *pogostost*

$$P_k = \lim_{N\to\infty} \frac{N_k}{N}\,. \tag{19.4}$$

Pri pošteni kocki, na primer, izmerimo v 1000 metih $P_3 = 0{,}17 \approx 1/6$ in enako za ostale izide. Pogostosti elementarnih izidov prikažemo s tabelo ali grafom – *frekvenčno porazdelitvijo* izidov.



Slika 19.2 Frekvenčna porazdelitev izidov pri metu poštene kocke. Vsak izid n se pojavlja z enako pogostostjo: porazdelitev je enakomerna.

Iz definicije je jasno, da mora za vsakršno frekvenčno porazdelitev veljati

$$\sum P_k = 1\,. \tag{19.5}$$

Rečemo, da so porazdelitve *normirane*.

Verjetnost izida

Čim večja je pogostost kakega izida v množici poskusov, tem bolj "verjetno" se nam zdi, da bo predstoječi posamični poskus pokazal ravno ta izid. Povedano izkoristimo za kvantitativno definicijo verjetnosti: *verjetnost* kakega izida pri posamičnem poskusu, to naj bo njegova relativna frekvenca v množici poskusov pri enakih "delovnih" pogojih. Pogostost se torej nanaša na množico poskusov, verjetnost pa na posamičen poskus. Izraz "verjetnost", kakor smo ga definirali in kakor ga hočemo uporabljati, ni nič drugega kot sinonim za izraz "pogostost". Verjetnosti so decimalna števila med 0 in 1.

19.4 Verjetnost sestavljenih izidov

Unija izidov

Kakšna je verjetnost, da pri metu kocke pade x_3 ali x_5? Da bomo bolj splošni, recimo: kakšna je verjetnost, da se v enem poskusu pokaže elementarni izid A ali elementarni izid B, torej vsaj eden izmed obeh? To je seveda tudi svojevrsten izid poskusa. Poimenujemo ga *unija* dveh elementarnih izidov ter ga označimo kot izid $(\mathrm{A}\cup\mathrm{B})$. Iz definicije verjetnosti neposredno sledi

$$P(\mathrm{A}\cup\mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B})\,. \tag{19.6}$$

Verjetnost, da se pri enem poskusu pokaže eden ali drugi od možnih elementarnih izidov, je enaka vsoti verjetnosti obeh posamičnih izidov. Da poštena kocka pokaže x_3 ali x_5, se zato zgodi z verjetnostjo $1/6 + 1/6 = 2/6$.

Pravilo o seštevanju verjetnosti ne velja le za dva elementarna izida, ampak tudi za več njih. Prav tako ne velja le za elementarne izide, temveč za kakršnekoli izide, ki se medsebojno izključujejo, to je, če se pokaže eden, se ne more hkrati pokazati še drugi. Dva takšna izključujoča se izida pri metu kocke sta, na primer: pade sodo število pik (x_2 ali x_4 ali x_6) in pade trojka (x_3). Verjetnost prvega izida je 1/2, verjetnost drugega je 1/6, in verjetnost njune unije, torej enega ali drugega, je $1/2 + 1/6 = 4/6$.

Presek izidov

Kakšna je verjetnost, da pri metu kocke pade x_3 in pri naslednjem metu x_5? Da bomo bolj splošni, recimo: kakšna je verjetnost, da se v prvem poskusu pokaže elementarni izid A in pri drugem poskusu elementarni izid B? To je tudi svojevrsten izid (dvojnega) poskusa. Poimenujemo ga *presek* obeh izidov ter ga označimo kot izid $(\mathrm{A} \cap \mathrm{B})$. Iz definicije verjetnosti neposredno sledi

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}) \cdot P(\mathrm{B})\,. \tag{19.7}$$

Verjetnost, da se pri prvem poskusu pokaže izid A in pri drugem izid B, je enaka produktu verjetnosti obeh posamičnih izidov. Seveda velja vse povedano tudi za več poskusov in za izide, ki niso elementarni. V vsakem primeru pa morajo biti poskusi medsebojno neodvisni, to je, izid drugega poskusa ne sme biti odvisen od izida prvega poskusa. Da poštena kocka pokaže prvič x_3 in druga x_5, se zato zgodi z verjetnostjo $1/6 \cdot 1/6 = 1/36$.

19.5 Binomska porazdelitev

Verjetnost, da pri metu kocke pade šestica, torej x_6, naj bo 1/6. Verjetnost, da ne pade šestica, pa je zato $1 - 1/6 = 5/6$. Zanima nas, kolikšne so verjetnosti, da v 5 metih pade šestica natanko 0-krat, 1-krat … 5-krat. Poskusi so sedaj petorke metov, opazovani izid pa število šestic, n, v vsaki petorki. Mečemo petorke v nedogled. Sproti štejemo, kolikokrat vsebujejo 0 šestic, 1 šestico in tako naprej. S tem so čedalje natančneje določene relativne frekvence P_n. Hočemo jih izračunati.

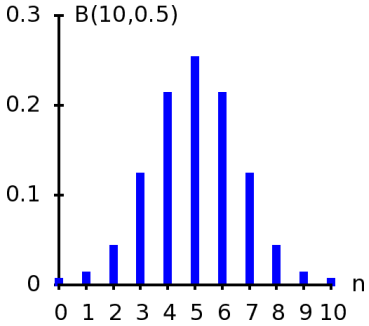
Število uspehov v vrsti poskusov

Bolj splošno lahko nalogo postavimo takole. Delamo take poskuse, ki imajo le dva izida, "uspeh" T in "neuspeh" F. Verjetnost za uspeh naj bo p in za neuspeh $1 - p = q$. Kakšna je verjetnost, da je v N poskusih natanko n uspešnih?

En način, na katerega se lahko pojavi $n = 2$ uspehov v $N = 5$ poskusih, je TTFFF. Verjetnost tega izida znaša $p \cdot p \cdot q \cdot q \cdot q = p^n\, q^{N-n}$. Vendar obstajajo še drugi načini, na primer FFFTT in TFFFT in še mnogi. Vsak izmed njih je enako verjeten, ker so zaporedni poskusi med seboj neodvisni. Verjetnosti vseh moramo sešteti. Koliko različnih N-teric pa pravzaprav lahko sestavimo iz n črk T in iz $(N - n)$ črk F? Toliko, kolikor je permutacij N elementov, od katerih je n enakih in $(N - n)$ tudi enakih: $N!/n!(N - n)!$. Iskana verjetnost je torej:

$$P(n) = \frac{N!}{n!(N-n)!}\, p^n (1-p)^{N-n} = \mathrm{B}_{N,p}(n)\,. \tag{19.8}$$

To je *binomska porazdelitev* (J. BERNOULLI). Pove nam, kakšna je verjetnost, da v N poskusih zadenemo natanko n uspešnih izidov, če je verjetnost takega izida pri posamičnem poskusu enaka p. Da v petih metih kocke pade natanko ena šestica, se torej zgodi z verjetnostjo 0,16.



Slika 19.3 Binomska porazdelitev. Prikazana je verjetnost, da v deseterici metov poštenega kovanca pade glava $0, 1, 2 \ldots 10$-krat.

Vsota verjetnosti vseh možnih izidov pri enem poskusu (N-terici metov) mora biti enaka ena, to je, porazdelitev verjetnosti mora biti normirana. Malo nas skrbi, ali to za izpeljano binomsko porazdelitev res drži. Eksplicitno zapisana vsota $\sum \mathrm{B}_{N,p}(n)$ znaša $C_N{}^0\, q^n + C_N{}^1\, pq^{n-1} + \ldots\, C_N{}^N\, p^n$. To pa ni nič drugega kot razviti binom $(q+p)^n$, torej $((1-p)+p)^n$, torej $1^n = 1$. Skrb je odveč, porazdelitev je normirana.

Slepo reševanje testov

Lep primer "uspešnega" poskusa je slepo reševanje šolskih testov. Učenec dobi 5 vprašanj. Ob vsakem so navedeni 3 odgovori in samo eden izmed njih je pravilen. Vsi odgovori se zdijo učencu enako verjetni, zato na slepo izbere enega. Verjetnost, da je prav uganil, je zato 1/3. Število uspehov, ki jih tako doseže, znaša od 0 do 5. Verjetnost, da doseže 4 ali 5 uspehov, je $\mathrm{B}_{5,1/3}(4) + \mathrm{B}_{5,1/3}(5) \approx 0{,}045$. Kaj takega se torej zgodi enkrat v $1/0{,}045 \approx 20$ testih.

Namesto da en učenec slepo opravi neskončno testov, si lahko mislimo neskončno učencev, ki na slepo opravijo en test. Frekvenčni porazdelitvi po rezultatih sta v obeh primerih enaki. Če je torej potrebnih ~ 20 testov, da en učenec slučajno doseže štiri ali pet točk, to slučajno uspe enemu izmed množice ~ 20 učencev.

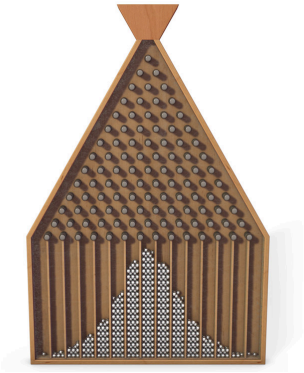
Še beseda o slepem izbiranju. Izbira enega izmed množice elementov, recimo enega izmed treh odgovorov, je slepa, če ima vsak element enako verjetnost, da je izbran. Dober način za to je naslednji: vse elemente oštevilčimo, številke zapišemo na listke in jih zapremo v čim bolj enake kroglice, vržemo kroglice v vrteč se boben ter čez nekaj časa z zavezanimi očmi potegnemo iz njega eno kroglico. Za prvo silo, če je elementov malo, zadostujejo kar prepognjeni listki in navaden klobuk. Da opisana načina res

zagotavljata enako verjetnost izbire, pa se na koncu koncev ne moremo prepričati nič drugače, kot da ju dejansko preizkusimo s štetjem izidov.

19.6 Vsota slučajnih izidov

Ožebljena deska

Na met in kotaljenje kocke učinkuje okolje z množico vplivov, ki jih ne poznamo in na katere je izid silno občutljiv. Majhna sprememba v začetnih in vmesnih pogojih, pa je rezultat že čisto drugačen. To nas navede na misel, da bi vpliv okolja na gibanje telesa lahko preučevali tudi tako, da bi po klancu spuščali kroglico, nanjo vplivali z gozdom zabitih žebljičkov, in gledali, kje na dnu bo pristala. Najpreprostejša je deska z N vrsticami žebljičkov, ki so med sabo razmaknjeni za premer kroglice, pri čemer je vsaka druga vrsta zamaknjena vstran za polovčno razdaljo med žebljički. To je ožebljena deska.



Slika 19.4 Ožebljena deska. Ilustracija deske, ki jo je uporabljal F. Galton. Spuščene kroglice se razvrstijo po binomski porazdelitvi. (Eterea Estudios)

Porazdelitev odmikov

Kroglico spustimo z vrha. Na prvi vrstici se odbije levo ali desno, na drugi prav tako in s cikcakanjem nadaljuje vse do dna. Verjetnost za odboj v desno naj bo vsakokrat p in za odboj v levo $q = 1 - p$. Ti dve verjetnosti sta ponavadi enaki. V N trkih opravi kroglica n korakov v desno in $N-n$ korakov v levo. Gibanje kroglice lahko torej opišemo kot N-kratni met kocke in štetje "ugodnih" izidov. Ugodni izid pri spuščanju kroglice je pač korak v (recimo) desno. Kolikokrat se bo kroglica premaknila v desno v N trkih, je torej opisano z binomsko porazdelitvijo $\mathrm{B}_{N,p}(n)$.

Neto premik v desno, m, je enak razliki premikov v desno in levo: $m = n - (N - n)$. Izrazimo n z m in ga vstavimo v binomsko porazdelitev, pri čemer izberemo še $p = q = 1/2$, pa dobimo:

$$\mathrm{B}_{N,1/2}(m) = \frac{N!}{[(N+m)/2]!\,[(N-m)/2]!}\,(\frac{1}{2})^N. \tag{19.9}$$

To je verjetnostna porazdelitev leg, ki jih doseže kroglica na dnu, oziroma delež kroglic, ki pristanejo v teh legah. Kadar izraza $N+m$ ali $N-m$ nista soda, bi morali računati faktorielo ulomnega števila. Kaj to pomeni, ne vemo in bo morda treba še primerno definirati. Zaenkrat bomo pri konkretnem računanju aproksimirali $(n+0{,}5)! \sim n!(n+1)/2$.

Dolga ožebljena deska

Če je ožebljena deska dolga, postane porazdelitev simetrično zvonasta. Kakšna je ta porazdelitev, ko raste N čez vse meje, pri čemer se omejimo še na področje $m \ll N$?

Faktoriele velikih števil so neznansko velike, zato porazdelitev najprej logaritmiramo. Nastane vsota logaritmov. Vsak člen oblike $\ln n!$ aproksimiramo z integralom: $\ln n! = \ln 1 + \ln 2 + \ldots \ln n \approx \int_1^n \ln x \,\mathrm{d}x$. Tak integral znaša $(x \ln x - x)\,|_1^n$, torej – ko zanemarimo še 1 v primeri z n – $\ln n! \approx n \ln n - n$. Nato pridobljene izraze $\ln(1 + m/N)$ aproksimiramo s kratko potenčno vrsto: $m/n - m^2/2N^2$. Dobimo $\ln \mathrm{B} \approx -m^2/2N$, torej

$$\mathrm{B}_{N,1/2}(m) \approx A \cdot \mathrm{e}^{-m^2/2N}. \tag{19.10}$$

Konstanto A smo pritaknili, ker sumimo, da smo zaradi številnih aproksimacij zapravili normiranost izhodiščne porazdelitve. To pomeni, da moramo to konstanto zdaj naknadno določiti iz pogoja normiranosti, torej $A = 1 / \int \exp(-m^2/2N)\,\mathrm{d}m$. S tem bo *normalna aproksimacija* k binomski porazdelitvi popolnoma določena.

Normalni integral

Kako izračunati *normalni integral* $I = \int \exp(-x^2)\,\mathrm{d}x$ med $-\infty$ in $+\infty$? Takole: $I^2 = \int \exp(-x^2)\,\mathrm{d}x \cdot \int \exp(-y^2)\,\mathrm{d}y = \iint \exp -(x^2 + y^2)\,\mathrm{d}x\mathrm{d}y$. To je ploskovni integral v kartezičnih koordinatah. Zapišemo ga v polarnih koordinatah $x^2 + y^2 = r^2$ in $\mathrm{d}x\mathrm{d}y = r\mathrm{d}r\mathrm{d}\varphi$, preoblikujemo $r\mathrm{d}r = {}^1\!/_2\,\mathrm{d}(r^2)$ in dobimo integral z navadno eksponentno funkcijo $I^2 = {}^1\!/_2 \iint \exp(-t)\,\mathrm{d}t\mathrm{d}\varphi$. Za meji med 0 in ∞ ter med 0 in 2π ga zlahka izračunamo in znaša π. Koren iz tega je torej iskani normalni integral:

$$\int_{-\infty}^{+\infty} \mathrm{e}^{-x^2}\,\mathrm{d}x = \sqrt{\pi}\,. \tag{19.11}$$

S tem je normalizacijska konstanta določena: $A = 1/\sqrt{(2\pi N)}$.

19.7 Normalna porazdelitev

Gostota verjetnosti

Ko z astrolabom določamo višino zvezde ob kulminaciji, se izmerki med seboj bolj ali manj razlikujejo. Če odmislimo *sistematične napake* – ko uporabimo nenatančen kotomer ali ko narobe odčitamo številko z njega ali ko celo merimo napačno zvezdo – preostane še množica *slučajnih napak* – zaradi nihanje astrolaba, migotanja ozračja in še kaj. Podobno se dogaja pri merjenju drugih količin. Izmerke takšne zvezne količine x razvrstimo v primerno široke razrede $x \pm \mathrm{d}x/2$ in preštejemo, koliko izmerkov $\mathrm{d}N(x \pm \mathrm{d}x/2)$ pade v vsakega. S tem je določena njihova frekvenčna porazdelitev

$$\frac{\mathrm{d}P}{\mathrm{d}x} = \lim_{N\to\infty} \frac{\mathrm{d}N(x \pm \mathrm{d}x/2)}{N} = p(x)\,, \tag{19.12}$$

ki je seveda normirana:

$$\int \mathrm{d}P = \int p(x)\,\mathrm{d}x = 1\,. \tag{19.13}$$

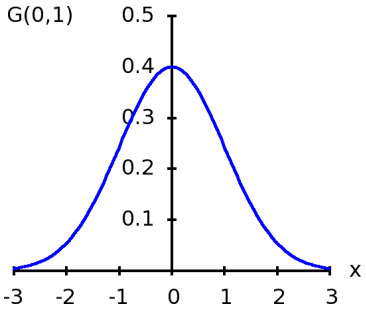
Pogledano z drugimi očmi: izmerek količine je slučajna spremenljivka in (limitna) frekvenčna porazdelitev izmerkov je njena gostota verjetnosti.

Normalna porazdelitev

Ko narišemo gostoto verjetnosti za izmerjene kulminacije ali kako drugo tovrstno količino, opazimo, da ima lepo zvonasto obliko, ki je na moč podobna normalni binomski aproksimaciji, le da je zvezna (19.10). Zato definiramo *normalno porazdelitev* kot (GAUSS)

$$\frac{\mathrm{d}P}{\mathrm{d}x} = \frac{1}{\sigma\sqrt{2\pi}} \cdot \mathrm{e}^{-(x-\mu)^2/2\sigma^2} = \mathrm{G}_{\mu,\sigma}(x)\,. \tag{19.14}$$

Parameter μ pove, kje leži vrh porazdelitve in parameter σ določa širino vrha. Kot kvadrat ga pišemo zato, da ima enake dimenzije kot slučajna spremenljivka. Sorazmernostna konstanta poskrbi za normiranost.



Slika 19.5 Normalna porazdelitev. Prikazana je porazdelitev s povprečjem 0 in deviacijo 1.

Dejstvo, da so kakšni izmerki porazdeljeni normalno, nam sporoča, da nanje vpliva – kakor na gibanje kroglice po žebljasti deski – množica med seboj neodvisnih in nasprotujočih si drobnih vplivov. Pravzaprav je normalna porazdelitev celo neke vrste zagotovilo, da izmerki niso obremenjeni s sistematičnimi, ampak zgolj s slučajnimi napakami.

Standardna porazdelitev

S porazdelitvijo verjetnosti po spremenljivki x je določena tudi porazdelitev po vsaki drugi, z njo povezani spremenljivki $z(x)$:

$$\frac{\mathrm{d}P}{\mathrm{d}z} = \frac{\mathrm{d}P}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}z}\,. \tag{19.15}$$

Če so izmerki x porazdeljeni kot $\mathrm{d}P/\mathrm{d}x = \mathrm{G}_{\mu,\sigma}(x)$, potem so ustrezajoči *normalizirani izmerki*

$$z = \frac{x-\mu}{\sigma} \tag{19.16}$$

porazdeljeni kot $\mathrm{d}P/\mathrm{d}z = (\mathrm{dG}/\mathrm{d}x)(\mathrm{d}x/\mathrm{d}z)$, torej takole:

$$\frac{\mathrm{d}P}{\mathrm{d}z} = \frac{1}{\sqrt{(2\pi)}} \cdot \mathrm{e}^{-z^2/2} = \mathrm{G}_{0,1}(z). \tag{19.17}$$

To je normalna porazdelitev z vrhom pri $\mu = 0$ in s širino $\sigma = 1$. Poimenujemo jo *standardna porazdelitev*. Verjetnost, da bo slučajni izmerek x ležal na intervalu med x_1 in x_2, je zato enaka

verjetnosti, da bo normalizirani izmerek z ležal na intervalu med $z_1 = (x_1 - \mu)/\sigma$ in $z_2 = (x_2 - \mu)/\sigma$. Ta verjetnost je enaka integralu $\mathrm{G}_{0,1}(z)$ med navedenima mejama. Za konkretno računanje potrebujemo še tabelirane vrednosti $\mathrm{G}_{0,1}(z)$ in njenega integrala

$$\int_0^z \mathrm{G}_{0,1}(z)\,\mathrm{d}z = \mathrm{erf}(z)\,. \tag{19.18}$$

Slednjega izračunamo z razvojem podintegralske funkcije $\exp t$, $t = -z^2/2$ v potenčno vrsto $1 + t + t^2/2! + \ldots$ in jo členoma integriramo:

$$\mathrm{erf}(z) = \frac{1}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{n!(2n+1)}\,. \tag{19.19}$$

Tako pridelamo tabelo

Tabela 19.1. Standardna porazdelitev in ploščina pod njo.

z	$\mathrm{G}_{0,1}$(z)	erf(z)
0.0	0,40	0,00
0.5	0,35	0,19
1.0	0,24	0,34
1.5	0,13	0,43
2.0	0,05	0,48
2.5	0,02	0,49
3.0	0,00	0,50

Verjetnost, da leži izmerek x znotraj intervala $\mu \pm \sigma$, je torej $2 \cdot 0{,}34 = 0{,}68$. Na intervalu $\pm 2\sigma$ leži z verjetnostjo $2 \cdot 0{,}48 = 0{,}95$. In na intervalu $\pm 3\,\sigma$ ga najdemo (skoraj) z gotovostjo $2 \cdot 0{,}50 = 1$.

19.8 Povprečje in varianca

Povprečje

Ko zaporedno zložimo N palic z dolžinami $l_1, l_2 \ldots l_N$, dobimo palico dolžine L. Enako dolgo sestavljeno palico dobimo tudi z N enakimi palicami dolžine $\bar{l}$, torej $N \cdot \bar{l} = \sum l_n$. S tem je definirana povprečna dolžina uporabljenih N palic: $\bar{l} = (1/N) \sum l_n$. Če je palic veliko in so nekatere med seboj enake, raje računamo takole: $\bar{l} = (1/N) \sum N_k\, l_k = \sum (N_k/N)\, l_k = \sum f_k\, l_k$. Keficienti f_k so relativne frekvence palic enake dolžine. Kar velja za palice in njihove dolžine, posplošimo za poljubno slučajno spremenljivko x: njeno *povprečno vrednost* v limitni množici poskusov, ko $f_k \to P_k$, definiramo kot $\langle x \rangle = \sum x_k P_k = \mathrm{Ave}(x)$. Če je spremenljivka zvezna, pa velja

$$\langle x \rangle = \int x\, p(x)\, \mathrm{d}x. \tag{19.20}$$

Vsota uteženih odmikov od povprečja je enaka nič: $\int (x - \langle x \rangle)\, \mathrm{d}P = \int x\, \mathrm{d}P - \langle x \rangle \int \mathrm{d}P = \langle x \rangle - \langle x \rangle = 0$.

Varianca in deviacija

Palice, iz katerih določamo povprečje, se med seboj bolj ali manj razlikujejo. Kolikšno je to razlikovanje, povemo s povprečnim

kvadratnim odmikom od povprečja: $s_l^2 = (1/N)\sum (l_n - \bar{l})^2$ oziroma $s_l^2 = \sum f_k (l_k - \bar{l})^2$. Kar velja za dolžino palic, posplošimo na poljubno slučajno spremenljivko: njeno *varianco* definiramo kot $\sigma_x^2 = \sum (x_k - \langle x \rangle)^2 P_k = \mathrm{Var}(x)$. Koren iz variance, σ_x, pa poimenujemo *deviacija*. Za zvezno spremenljivko velja:

$$\sigma_x^2 = \int (x - \langle x \rangle)^2 p(x)\,\mathrm{d}x\,. \tag{19.21}$$

Integral lahko preoblikujemo: kvadriramo podintegralski binom, integriramo dobljene člene in pridelamo izraz

$$\sigma_x^2 = \int x^2 p(x)\,\mathrm{d}x - (\int x\, p(x)\,\mathrm{d}x)^2 = \langle x^2 \rangle - \langle x \rangle^2\,. \tag{19.22}$$

Izračun povprečij in varianc

Če so porazdelitve podane s tabelo, računamo njihova povprečja in variance s konkretnimi številskimi vrednostmi. Če so podane z enačbo, pa lahko računamo s simboli. Izračunajmo povprečja in variance tistih porazdelitev, ki smo jih že spoznali!

Za enakomerno diskretno porazdelitev (pošteno kocko) velja $\langle x \rangle = \sum n \cdot (1/6) = 3{,}5$ in $\sigma_x^2 = \sum n^2 \cdot (1/6) - (3{,}5)^2 = (1{,}7)^2$. Na interval $\langle x \rangle \pm \sigma_x$ padejo vrednosti 2, 3, 4 in 5, to je, 2/3 vseh vrednosti.

Za binomsko porazdelitev že poznamo njeno vsoto: $\sum C_N^n\, p^n\, q^{N-n} = (p+q)^N$. Če bi bil vsak člen vsote pomnožen s faktorjem n, bi nastala vsota opisovala povprečje. Kako pridelati faktorje n? Levo in desno stran odvajamo na p in nato množimo s p. Na levi nastane povprečje $\langle x \rangle = \sum n\, C_N^n\, p^n\, q^{N-n}$ in na desni izraz $np\,(p+q)^{N-1}$. Ko v njem upoševamo $q = 1 - p$, najdemo $\langle x \rangle = Np$. Podobno izračunamo varianco – izhodiščno enačbo dvakrat odvajamo na p in nato pomnožimo s p^2. Tako dobimo $\sigma_x^2 = Npq$.

Pri računanju povprečja in variance normalne porazdelitve moramo izračunati integrala oblike $\int x \exp(-x^2)\,\mathrm{d}x$ in $\int x^2 \exp(-x^2)\,\mathrm{d}x$. Prvega izračunamo tako, da spravimo x pod diferencial, s čimer prevedemo integral v lahko rešljivo obliko $\int \exp(-t)\,\mathrm{d}t$. Drugega pa se lotimo po delih: $u = x$, $\mathrm{d}v = x \exp(-x^2)\,\mathrm{d}x$ in ga s tem prevedemo ne integral za povprečje. Dobimo $\langle x \rangle = \mu$ in $\sigma_x^2 = \sigma^2$.

Katerokoli porazdelitev, ki ima povprečje $\langle x \rangle$ in varianco σ_x^2, lahko aproksimiramo z normalno porazdelitvijo, ki ima isto povprečje in varianco. Ujemanje je bolj ali manj dobro. Normalna aproksimacija enakomerne porazdelitve je prav slaba, binomske pa naravnost odlična, če je le njen parameter N dovolj velik. Nekaj konkretnih grafov pokaže, da je ujemanje precej dobro že pri $N = 10$.

19.9 Večdimenzijske porazdelitve

Pri nadaljnji raziskavi bo očitno nerodno uporabljati dve različni pisavi, eno za diskretne primere in drugo za zvezna primere. Odločimo se, da bomo uporabljali le pisavo za zvezno

spremenljivko, ki pa jo v bomo primeru diskretnosti razumeli takole: $p(x)\mathrm{d}x \to P_k$ in $\int p(x)\mathrm{d}x \to \sum P_k$.

Dve spremenljivki

Pri streljanju s puško v tarčo je lega zadetka slučajna spremenljivka.

Slika 19.6 Tarča. Lega zadetka je slučajna spremenljivka. (Anon)

Vsak zadetek ima svoj vodoravni odmik x in navpični odmik y od središča tarče. Gostoto verjetnosti za zadetek okrog točke (x,y), to je na intervalu $(x \pm \mathrm{d}x/2, y \pm \mathrm{d}y/2)$, definiramo s številom strelov $\mathrm{d}N$ v ta interval, deljenim s številom vseh strelov N:

$$\frac{\mathrm{d}^2 P}{\mathrm{d}x\,\mathrm{d}y} = \lim_{N\to\infty} \frac{\mathrm{d}N(x \pm \mathrm{d}x/2, y \pm \mathrm{d}y/2)}{N} = p(x,y). \tag{19.23}$$

Predstavljamo si jo kot ploskev oziroma kot hrib, ki je ponekod bolj, drugod manj visok. Višina hriba na nekem mestu pove, kakšna je tamkajšnja pogostost oziroma verjetnost zadetkov.

Robne verjetnosti

Verjetnost za vodoravni izid okrog x, neodvisno od tega, kakšen je navpični izid, je vsota

$$\frac{\mathrm{d}P}{\mathrm{d}x} = \int p(x,y)\,\mathrm{d}y = u(x)\,. \tag{19.24}$$

Predstavljamo si, da smo ves hrib stlačili na vodoravno os, vzdolž katere se je naredil kumulativni profil $u(x)$. Podobno velja tudi za tlačenje hriba na navpično os, ko nastane kumulativni profil $v(y)$.

Pogojne verjetnosti

Kolikšna pa je verjetnost za vodoravni izid okrog x pri pogoju, da je navpični izid okrog y? Vzdolž ozkega vodoravnega pasu okrog $y = \mathrm{const}$ definiramo verjetnost

$$\frac{\mathrm{d}P}{\mathrm{d}x}\Big|_y = \lim_{N\to\infty} \frac{\mathrm{d}N(x \pm \mathrm{d}x/2)}{N(y \pm \mathrm{d}y/2)} = p(x\,|\,y). \tag{19.25}$$

Rekli bomo, da je to *pogojna verjetnost* za izid okrog x glede na izid okrog y. Predstavljamo si jo kot profil hriba vzdolž vodoravnega prereza. Seveda velja podobno tudi za pogojne verjetnosti vzdolž navpičnih pasov, $p(y\,|\,x)$. Iz definicij verjetnosti, robne verjetnosti in pogojne verjetnosti sledi

$$p(x,y) = u(x)\,v(y|x)\,. \tag{19.26}$$

Res. Verjetnost za strel okrog (x,y) je enaka robni verjetnosti za strel okrog x, pomnoženi z ustrezno pogojno verjetnostjo za strel

okrog y. Kadar je slučajna spremenljivka y neodvisna od x, je njena pogojna verjetnost $v(y|x)$ kar enaka "nepogojni" verjetnosti $v(y)$ in velja že znano produktno pravilo (19.7)

$$p(x,y) = u(x)\,v(y)\,. \tag{19.27}$$

Dober primer je streljanje v tarčo, če nastane gostota $\exp(-r^2)$, to je $\exp(-x^2-y^2)$, torej $\exp(-x^2)\cdot\exp(-y^2)$. Strelca zanaša v levo in desno enako, neodvisno od tega, kako ga zanaša gor in dol, in obratno.

19.10 Soodvisnost spremenljivk

Povprečje in varianca

Za vsako spremenljivko posebej lahko definiramo njeno povprečje in varianco. Za spremenljivko x tako velja:

$$\begin{aligned}\langle x\rangle &= \int\!\!\int x\,p(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ \sigma_x{}^2 &= \int\!\!\int (x-\langle x\rangle)^2\,p(x,y)\,\mathrm{d}x\,\mathrm{d}y\,.\end{aligned} \tag{19.28}$$

Očitno sta to povprečje in varianca robne verjetnosti: $\langle x\rangle = \int x\,u(x)\,\mathrm{d}x$ in $\sigma_x{}^2 = \int (x-\langle x\rangle)^2\,u(x)\,\mathrm{d}x$. Podobno velja za spremenljivko y.

Kovarianca in korelacija

Sama se ponuja še mešana količina

$$\sigma_{xy} = \int\!\!\int (x-\langle x\rangle)(y-\langle y\rangle)\,p(x,y)\,\mathrm{d}x\,\mathrm{d}y\,. \tag{19.29}$$

Poimenujemo jo *kovarianca*. Pričakujemo, da na nek način pove, kako močno sta spremenljivki med seboj odvisni. Preverimo to domnevo! Če sta spremenljivki neodvisni, torej če $p(x)=u(x)v(y)$, se kovariantni integral zapiše kot produkt dveh integralov, od katerih je vsak enak nič, torej je tudi kovarianca enaka nič. Če sta spremenljivki natanko sorazmerni, torej $y=kx$, so odmiki od povprečij maksimalni in koviariantni integral se reducira v $k\sigma_x{}^2$ oziroma v $(1/k)\sigma_y{}^2$. Domneva je torej potrjena. Zato je smiselno definirati

$$r = \frac{\sigma_{xy}}{\sigma_x\sigma_y}\,, \tag{19.30}$$

to je *korelacijski koeficient* dveh spremenljivk. Koeficient očitno leži med vrednostima -1 in 1. Čim večja je njegova absolutna vrednost, tem tesnejša je medsebojna odvisnost spremenljivk.

19.11 Vzorčenje in statistika

Populacija in vzorci

Povprečje in varianco smo definirali za neskončno veliko množico poskusov oziroma opazovanj oziroma meritev, to je na neskončni (ali zelo veliki) *populaciji*. Rekli bomo, da sta to *populacijska parametra*. Določimo ju pa seveda lahko le iz končnega *vzorca*; tedaj jima bomo rekli *vzorčni statistiki*.

Vzorčne statistike so seveda le približek k ustreznim populacijskim parametrom. Če je vzorec velik in slepo izbran, pričakujemo, da je ujemanje dobro. Pojavi se vprašanje, kako

točne so takšne ocene, to je, kolikšne napake pri tem zagrešimo. Poskusimo to narediti za povprečje!

Ko opravimo N poskusov in zabeležimo njihove izide, s tem iz neskončne populacije poskusov izberemo končni vzorec. Za ta vzorec izračunamo povprečje $\bar{x}$. Pri kakem drugem vzorcu bi dobili drugačno povprečje. Mislimo si, da vzorčenje kar naprej ponavljamo. Dobimo neskončno populacijo povprečij. Kakšna je njihova povprečna vrednost $\langle \bar{x} \rangle$? In kakšna je njihova varianca $\sigma_{\bar{x}}^2$?

Povprečje povprečij

Na izmerjene vzorčne vrednosti $x_1 \ldots x_N$ lahko pogledamo kot na uresničitev N slučajnih, med seboj neodvisnih spremenljivk $X_1 \ldots X_N$ iz osnovne populacije. Vse so porazdeljene tako, kot osnovna spremenljivka X. Spremenljivka X_1 je pri vzorčenju pač pokazala vrednost x_1, pri drugem vzorcu bi pa pokazala kaj drugega. Podobno velja za druge spremenljivke. Izmerjeno povprečje $\bar{x}$ pa je potem uresničitev slučajne spremenljivke $\bar{X} = (1/N) \sum X_n$.

Kakšno je torej povprečje vzorčnih povprečij $\langle \bar{X} \rangle = \mathrm{Ave}(X_1 + \ldots X_N)/N)$? Izpostavimo faktor $1/N$ izven povprečja; povprečje vsote je vsota povprečij; povprečje X_n je povprečje X; in dobimo:

$$\langle \bar{X} \rangle = \langle X \rangle \,. \tag{19.31}$$

Povprečje vzorčnih povprečij je torej enako populacijskemu povprečju. To je dobro.

Varianca povprečij

In kakšna je varianca vzorčnih povprečij $\sigma_{\bar{X}}^2 = \mathrm{Var}((X_1 + \ldots X_N)/N)$? Izpostavimo faktor $1/N$ izven variance, pri čemer postane $(1/N)^2$; varianca vsote je vsota varianc; varianca X_n je varianca X; in dobimo:

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{N} \,. \tag{19.32}$$

Vzorčna povprečja se torej stiskajo okrog populacijskega povprečja z N-krat manjšo varianco, kot je varianca posamičnih spremenljivk. Tudi to je dobro.

Porazdelitev povprečij

Vzorčno povprečje je (normirana) vsota N neodvisnih slučajnih spremenljivk z isto porazdelitvijo. To močno spominja na pot kroglice po ožlebljeni deski: ena pot, ki jo kroglica ubere, je en vzorec z N spremenljivkami, njihova vsota pa je končni odmik kroglice na dnu. Spremenljivke so "binomske", imajo samo dva izida. Vsote velikega števila binomskih spremenljivk se torej porazdelijo normalno. Morda velja to tudi za vsote velikega števila "nebinomskih" spremenljivk? Domnevamo torej

$$\frac{\mathrm{d}P}{\mathrm{d}\bar{X}} \propto \exp\left[-\frac{1}{2}\left(\frac{\bar{X} - \langle \bar{X} \rangle}{\sigma_{\bar{X}}}\right)^2\right] . \tag{19.33}$$

Ni videti lahke poti, da bi z doslej pridobljenim znanjem domnevo dokazali. Pa nič hudega: saj jo lahko utrdimo eksperimentalno. Mečemo pošteno kocko. Na stranice v mislih napišemo 1, 2, 3, 3, 4, 5. Verjetnostna porazdelitev izidov je zato $P(1) = 1/6$, $P(2) = 1/6$, $P(3) = 1/3$, $P(4) = 1/6$ in $P(5) = 1/6$, torej ima $\langle x \rangle = 3{,}0$ in $\sigma_x = 1{,}7$. Kocko vržemo 10-krat in dobimo prvi vzorec ter njegovo povprečje (nekje med 1,0 in 5,0). To ponovimo stokrat. Dobljenih sto povprečij porazdelimo po primerno širokih razredih. Porazdelitev se kar dobro prilega pričakovani normalni z $\mu = 3{,}0$ in $\sigma = 1{,}7/\sqrt{10} = 0{,}5$. Daljši vzorci in številčnejše ponovitve pokažejo še boljše prileganje. Seveda lahko kockine stranice kakorkoli oštevilčimo. Bolj kot je osnovna porazdelitev različna od normalne, daljše vzorce potrebujemo, da je njihova povprečna vrednost zadovoljivo normalno porazdeljena.

19.12 Merjenje in merske napake

Natančnost meritev

Povedano uporabimo za oceno merskih napak. Večkratna meritev kakšne količine, recimo dolžine mize, je namreč slučajno vzorčenje. Merjena dolžina je slučajna spremenljivka. Izmerjeno povprečje in varianca pa sta dve statistiki, iz katerih sklepamo na "pravo" dolžino mize. Ocenimo $\bar{x} \approx \langle x \rangle \pm \sigma_x / \sqrt{N}$. Neznano populacijsko deviacijo σ_x aproksimiramo kar z znano vzorčno deviacijo s_x, pa z nekaj drznosti zapišemo

$$\langle x \rangle \approx \bar{x} \pm \frac{s_x}{\sqrt{N}} \,. \tag{19.34}$$

Kadar je izmerkov malo, se ni treba mučiti z izračunom s_x. Kar na oko ocenimo, kakšen je interval okrog povprečja, v katerega pade 2/3 izmerkov, in zapišemo $\langle x \rangle \approx \bar{x} \pm \mathrm{d}x = \bar{x}(1 \pm \mathrm{d}x/\bar{x})$. Količino $\mathrm{d}x$ poimenujemo *absolutna napaka* in $\mathrm{d}x/\bar{x}$ *relativna napaka*.

Izboljšanje natančnosti

Čim več je meritev, tem manjša odstopanja njihovega povprečja od prave vrednosti pričakujemo. Večkratno merjenje je torej dober način, da izboljšamo natančnost izmerka. Žal pa se z naraščanjem N povečuje $\sqrt{N}$ le počasi. Če hočemo natančnost povečati za faktor 10, moramo povečati število meritev za faktor 100. Pri tem pa niti ne zmanjšujemo sistematičnih napak.

Širjenje napak

Če je kakšna količina obremenjena z napako, in to je zmeraj, so tudi njene funkcije obremenjene z napakami. Rečemo, da se napake podedujejo oziroma se širijo. Kako to gre?

Na napako funkcije lahko pogledamo kot na njen diferencial. Pri funkciji ene spremenljivke je to navadni diferencial in pri funkciji več spremenljivk imamo opravka s totalnim diferencialom. Seveda pa moramo upoštevati, da so takšni diferenciali lahko pozitivni ali negativni. Tako z diferenciranjem dobimo naslednja pravila.

$$
\begin{aligned}
&u = cx \Longrightarrow \mathrm{d}u = |c|\,\mathrm{d}x \\
&u = x \pm y \Longrightarrow \mathrm{d}u = \mathrm{d}x + \mathrm{d}y \\
&u = xy \Longrightarrow \frac{\mathrm{d}u}{|u|} = \frac{\mathrm{d}x}{|x|} + \frac{\mathrm{d}y}{|y|} \\
&u = \frac{x}{y} \Longrightarrow \frac{\mathrm{d}u}{|u|} = \frac{\mathrm{d}x}{|x|} + \frac{\mathrm{d}y}{|y|} \\
&u = x^n \Longrightarrow \frac{\mathrm{d}u}{|u|} = |n|\,\frac{\mathrm{d}x}{|x|} \\
&u = u(x) \Longrightarrow \mathrm{d}u = |u'|\,\mathrm{d}x \\
&u = u(x,y) \Longrightarrow \mathrm{d}u^2 = (u_x\,\mathrm{d}x)^2 + (u_y\,\mathrm{d}y)^2\,.
\end{aligned}
\tag{19.35}
$$

Napaka vsote ali razlike je vsota napak posameznih členov. Relativna napaka produkta ali kvocienta pa je vsota relativnih napak posameznih faktorjev. Zlasti je nevarno takrat, kadar naletimo na razliko dveh približno enakih členov. Tedaj je relativna napaka lahko ogromna. Računanje odvodov je včasih zoprno. V takem primeru lahko ocenimo kar $\mathrm{d}u \approx u(x+\mathrm{d}x) - u(x)$ oziroma $\mathrm{d}u \approx u(x+\mathrm{d}x, y+\mathrm{d}y) - u(x,y)$ za primerno izbrane neodvisne diferenciale.

19.13 Intervalno ocenjevanje

Ko rečemo $\bar{x} = \mu \pm \sigma/\sqrt{N}$, pravzaprav pravimo, da leži μ nekje na intervalu $[\bar{x} - \sigma/\sqrt{N}, \bar{x} + \sigma/\sqrt{N}]$ z verjetnostjo 0,68 in izven tega intervala z verjetnostjo 0,32. Oceno za μ pa lahko podamo bolj na splošno takole: leži na intervalu $[\bar{x} - x_\alpha, \bar{x} + x_\alpha]$ z verjetnostjo α, na primer 0.95. Kakšna je povezava med x_α in α?

Verjetnostni interval

Vemo tole. Če je $\bar{X}$ porazdeljen normalno kot $\mathrm{G}_{\mu,\sigma/\sqrt{N}}$, potem je $Z = (\bar{X} - \mu)/(\sigma/\sqrt{N})$ porazdeljena normalno kot $\mathrm{G}_{0,1}$. To pomeni, da je

$$
\begin{aligned}
&P(-z_\alpha \le Z \le +z_\alpha) = P(\bar{X} - x_\alpha \le \mu \le \bar{X} + x_\alpha) = 2\,\mathrm{erf}(z_\alpha) = \alpha \\
&x_\alpha = z_\alpha \sigma/\sqrt{N}\,.
\end{aligned}
\tag{19.36}
$$

Za vsako izbrano verjetnost α lahko izračunamo pripadajočo vrednost x_α. Verjetnosti 0,68, na primer, odgovarja $z_\alpha = 1$, torej $x_\alpha = \sigma/\sqrt{N}$, kakor tudi mora biti. Verjetnosti 0,95 pa odgovarja 2-krat tolikšen interval. Če hočemo v več primerih uloviti srednjo vrednost μ, moramo pač razširiti lovilno past.

Ocena intervala

Za izračun x_α moramo poznati deviacijo populacije. Te ponavadi ne poznamo, zato jo aproksimiramo kar z deviacijo vzorca. Širino intervala, ki pri 95 % vzorcev vsebuje neznano povprečje μ, torej določimo takole. Potegnemo vzorec dolžine N, iz njega izračunamo $\bar{x}$ in s_x ter izračunamo $x_{0.95} = 2s_x/\sqrt{N}$. S tem je interval izračunan. Če ga hočemo prepoloviti, potrebujemo štirikrat večji vzorec.

Verjetnost, da ocenjeni *interval zaupanja* dejansko pokrije neznano pravo povprečje, znaša α. Rečemo, da je to *stopnja zaupanja*. Seveda pa tvegamo, da povprečje leži izven intervala.

Verjetnost, da se to zgodi, znaša $1-\alpha$. Rečemo, da je to *stopnja tveganja*.

19.14 Preizkušanje domnev

Domneva o povprečju

Vojaški zdravnik trdi, da je povprečna višina v populaciji vojakov $\langle x \rangle = a$. To domnevo hočemo preveriti. Če domneva drži, vemo, da je vzorčna statistika $Z = (\bar{X} - a)/(\sigma_x/\sqrt{N})$ porazdeljena standardno kot $\mathrm{G}_{0,1}(Z)$. Ker ne poznamo populacijske deviacije, jo aproksimiramo z vzorčno deviacijo in dobimo statistiko $T = (\bar{X} - a)/(S_x/\sqrt{N})$. Pričakujemo, da je tudi ona porazdeljena približno kot $\mathrm{G}_{0,1}(T)$. To pomeni, da je na intervalu $[-t_\alpha, +t_\alpha] = [-2, +2]$ pričakovati $\alpha = 95\,\%$ uresničitev te statistike. Da pade uresničitev izven intervala, pa pričakujemo le v 5 % vzorcev. Iz populacije torej na slepo potegnemo vzorec N vojakov in izračunamo $\bar{x}$, s_x ter iz obojega t. Če pade t znotraj postavljenega intervala, nimamo kaj reči. Če pa pade t izven tega intervala, lahko to razlagamo na dva načina: — domneva je sicer pravilna, a smo imeli tako nesrečno roko, da smo naleteli na enega izmed tistih 5 % vzorcev; — domneva je vsekekor nepravilna. Katero izmed obeh razlag izbrati? Odločimo se, da je bolj verjetna druga razlaga in domnevo zavrnemo.

Dve vrsti napak

S preizkušanjem domnev torej ne sprejemamo, ampak jih zgolj – bolj ali manj utemeljeno – zavračamo. Očitno lahko pri tem naredimo dve vrsti napak: domneve ne zavrnemo, čeravno je nepravilna, ali pa domnevo zavrnemo, čeravno je pravilna. Kadar ima zavračanje domneve hude posledice, hočemo biti nadvse gotovi, da jo zavračamo utemeljeno. Takrat gledamo interval $[-3, +3]$ in ustrezno verjetnost 99,8 %.

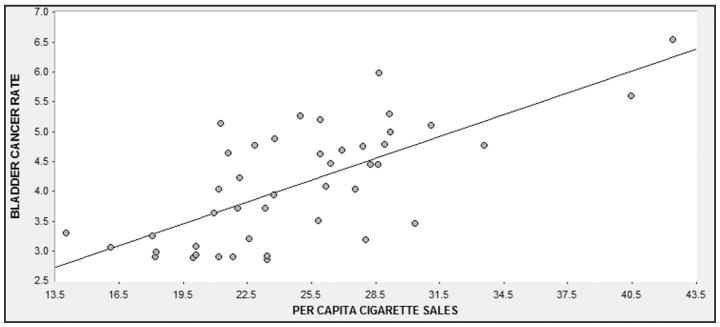
Ko zavračamo domnevo, moramo vsekekor povedati, pri kakšni stopnji tveganja $1-\alpha$ to počnemo. Tako rečemo, da smo domnevo zavrnili pri stopnji tveganja 5 %, oziroma da se vzorčni podatki statistično značilno razlikujejo od domneve pri tej stopnji tveganja. Stopnja tveganja pove, kolikšna je verjetnost, da smo domnevo zavrnili, čeravno je pravilna.

Druge domneve

Domnevamo, da lahko na podoben način zavračamo najrazličnejše domneve o populacijah, na primer: varianca porazdelitve je enaka neki vrednosti; povprečji dveh porazdelitev sta enaki; varianci dveh porazdelitev sta enaki; porazdelitvi sta enaki; in še kaj. Postopek je vedno enak: postaviti moramo ustrezno cenilko in zanjo določiti porazdelitev. Potem pogledamo, kako verjetna je dejanska uresničitev cenilke in se glede na to odločamo. Vse to je seveda lažje reči kot narediti. Podrobnejšo obravnavo zato prepustimo tistim, ki to potrebujejo (FISCHER).

19.15 Regresijska analiza

Soodvisnost dveh spremenljivk, tabeliranih v N parih (x_n, y_n) lahko aproksimiramo s premico, ki se jima "najbolj prilega". Najboljše prileganje definiramo takole: vsota kvadratov odmikov ene spremenljivke od premice naj bo minimalna. Minimiziramo lahko odmike y_n ali x_n; v splošnem se dobljeni premici razlikujeta. Najbolje je minimizirati odmike tiste spremenljivke, ki ima večjo deviacijo. Naj bo to spremenljivka y. Zaradi preprostosti še privzamemo, da so deviacije spremenljivke x enake nič.



Slika 19.7 Povezava med kajenjem in rakom. Za 44 ameriških držav je bilo določeno, koliko cigaret na prebivalca je bilo prodanih v letu 1960 in koliko smrti na 100 tisoč prebivalcev zaradi raka na mehurju je bilo zabeleženih v istem letu. (Fraumeni, 1968)

Določitev koeficientov

Iščemo torej funkcijo

$$y^* = A + Bx \tag{19.37}$$

tako, da bo $\sum (y^*_n - y_n)^2 = \sum (A + Bx_n - y_n)^2 = Q(A, B)$ minimalen. Postavimo $\partial Q/\partial A = 0$ in $\partial Q/\partial B = 0$, s čimer pridelamo dve linearni enačbi z dvema neznankama A in B: $AN + B\sum x_n = \sum y_n$ in $A\sum x_n + B\sum {x_n}^2 = \sum x_n y_n$. Iz enačb izračunamo obe neznanki in s tem je regresijska premica določena (GAUSS):

$$A = \frac{(\sum {x_n}^2)(\sum y_n) - (\sum x_n)(\sum x_n y_n)}{\Delta} \tag{19.38}$$

$$B = \frac{N(\sum x_n y_n) - (\sum x_n)(\sum y_n)}{\Delta}$$

$$\Delta = N(\sum {x_n}^2) - (\sum x_n)^2.$$

Ocena napak

Vzorčne vrednosti y_n imamo lahko za uresničitev slučajnih spremenljivk Y_n. Predpostavimo, da je vsaka izmed teh spremenljivk porazdeljena normalno okrog svoje srednje vrednosti $A + Bx_n$ z isto "lokalno" deviacijo σ. Zato so vse spremenljivke $Y_n - A - Bx_n$ porazdeljene normalno kot $G_{0,\sigma}$. Iz tega sklepamo, da je dobra ocena za lokalne deviacije kar enaka "globalni" deviaciji

$$s_y{}^2 = \frac{1}{N}\sum (y_n - A - Bx_n)^2 . \tag{19.39}$$

Parametra A in B sta čisti funkciji izmerkov $y_1 \ldots y_N$. Zato sta njuni deviaciji oz. napaki s_A in s_B določeni kar z deviacijami oz. napakami s_y slednjih. V obrazec za širjenje napak $s_A{}^2 = \sum (\partial A/\partial y_n \cdot s_y)^2$ vstavimo $\partial A/\partial y_n = [(\sum x_n{}^2) - x_n(\sum x_n)]/\Delta$ in dobimo, po nekaj računanja,

$$s_A{}^2 = s_y{}^2 \sum x_n{}^2 / \Delta \qquad (19.40)$$
$$s_B{}^2 = s_y{}^2 N / \Delta \,.$$

Podobno obravnavamo tudi linearno regresijo več spremenljivk. Kogar to veseli, pa se lahko loti celo nelinearne regresije.

19.16 Statistično zavajanje

Pravijo, da obstajajo tri vrste laži: navadna laž, huda laž in statistika. Nedvomno je res, da je statistika močno orodje za raziskavo množice podatkov, če jo seveda prav uporabljamo. Je pa tudi res, da se jo da zlorabiti na najrazličnejše načine. Pogosto to počno politiki in prodajalci. Kakšni so njihovi glavni načini zavajanja?

Majhen vzorec

Osnova statistike je vzorčenje. Vzorec mora biti dovolj velik, da iz njega lahko karkoli sklepamo. Beremo recimo, da se 33,3 % študentk na univerzi N. N. poroči s svojimi profesorji. Natančne številke in decimalna mesta nas prepričujejo, da raziskovalec ve, o čem govori. Surove številke pa govorijo drugače: v obdobju raziskave so bile na univerzi vpisane tri študentke, od katerih se je ena poročila s profesorjem.

Neslučajen vzorec

Vzorec mora biti tudi slučajen. Ko anketiramo ljudi, mora imeti vsak človek enako verjetnost, da ga izberemo. Beremo recimo, da 73 % Slovencev nasprotuje smrtni kazni. Vprašamo se: katerih Slovencev? Pokaže se, da je raziskavo naredil levičarski časopis N. N. preko vprašalnikov, ki jih je kar priložil časopisu. Ta časopis kupujejo pretežno levičarji in ti imajo bolj odklonilen odnos do smrtne kazni kot desničarji. Sklepanje na celotno populacijo je povsem neutemeljeno.

Golo povprečje

Povprečje nič ne pove o razpršenosti izmerkov okrog njega. Podjetje N. N. na primer objavi, da znaša povprečna mesečna plača njihovega delavca solidnih 3000 dolarjev. Lepo in prav, dokler ne odkrijemo, da je v podjetju zaposlenih 9 delavcev in en direktor. Direktor ima 21.000 dolarjev plače in delavci po mizernih 1000 dolarjev. Skoraj vsakdo je pod navedenim povprečjem!

Korelacija kot vzrok

Korelacija ne pomeni vzročne odvisnosti. Študentje, ki kadijo, imajo nižje ocene. To je verodostojno statistično dokazano. Torej kajenje povzroča slabe ocene? Morda celo otopi možgane? Nič od tega: če gresta kajenje in slabe ocene skupaj, to še ne pomeni, da kajenje povzroča slabe ocene. Morda je ravno obratno: slabe ocene silijo študente h kajenju. Ali pa nobeno ne povzroča

drugega, marveč je oboje posledica kakega tretjega vzroka. Je morda tako, da družabni ljudje, ki ne jemljejo preveč resno knjig, hkrati tudi kadijo več?

Obrezani grafi

Kako cene rastejo, najlepše pokažemo z grafom. Recimo, da kakšna cena v desetih letih naraste od 100 na 110 dolarjev. Na grafu z višino 5 cm, ki ima navpično os oštevilčeno od 0 do 120, je rast cene zelo položna krivulja. Morda nam to ni všeč? Odrežimo spodnji in zgornji del grafa (z izgovorom, da sta itak prazna) ter prikažimo zgolj navpični interval med 100 in 110 dolarji, seveda raztegnjen na isto višino. Mnogo bolje! Graf je sedaj zelo strma krivulja, ki kar kriči, kakšen hud porast cen se je zgodil. Nič ni bilo ponarejenega – razen vtisa, ki ga graf zapusti. Podobno lahko polepšamo tudi druge vrste grafov.

Obramba

Kako si pomagamo, da nas takšne "statistike" in sklepi iz njih ne zavedejo? Tako, da odgovorimo na nekaj vprašanj. Kdo to pravi? Kako to ve? Kaj vse manjka (velikost vzorca, način vzorčenja, povprečje brez deviacije, testiranje domnev brez stopnje tveganja, korelacijski parametri brez ocenjenih napak, grafi brez meril)? Ali je vse skupaj smiselno? Nikoli pa tudi ne smemo pozabiti, da je statistika vredna zgolj toliko, kot so verodostojni podatki, na katerih sloni. □