0. Kazalo

1. Osnovni pojmi verjetnostnega računa

- **poskus**, **dogodek** in **verjetnost** dogodkov
- **kompleks pogojev** (množica pojavov, ki vedno nastopajo skupaj)
- **gotov** dogodek *G* (verjetnost=1), **nemogoč** dogodek *N* (verjetnost=0), **slučajen** dogodek (0<verjetnost<1)

$f(A) = \frac{k}{n}$

$f(A)$	relativna frekvenca dogodka A
k	frekvenca dogodka A (število uspehov)
n	število ponovitev poskusa

Statistična definicija verjetnosti:
Verjetnost *P(A)* v danem poskusu je število, pri katerem se stabilizira relativna frekvenca dogodka *A* v velikem številu ponovitev tega poskusa.

2. Dogodki

- dogodek *A* je **način** dogodka *B* ($A \subset B$ ali $B \supset A$), če se vedno z dogodkom *A* zgodi tudi dogodek *B*
- če je $A \subset B$ in $B \subset A$, se zgodita *A* in *B* vedno hkrati, torej sta **enaka** ($A = B$)
- **vsota** dogodkov ($A + B$ ali $A \cup B$) je dogodek, da se zgodi vsaj eden od obeh
- **produkt** dogodkov (AB ali $A \cap B$) je dogodek, da se zgodita oba hkrati
- **nezdružljiva** dogodka sta dogodka, ki se ne moreta zgoditi hkrati, njun produkt je nemogoč dogodek; **paroma nezdružljivi** dogodki so dogodki, od katerih se niti dva ne moreta zgoditi hkrati
- $\overline{A}$ je **negacija** dogodka *A*; ta dva dogodka sta si **nasprotna**, ker se v vsaki ponovitvi poskusa zgodi natanko eden od njiju ($A\overline{A} = N$ in $A + \overline{A} = G$)
- **popoln sistem** dogodkov je množica dogodkov, od katerih se v vsaki ponovitvi poskusa zgodi natanko eden
- če lahko dogodek *A* izrazimo kot vsoto vsaj dveh mogočih nezdružljivih dogodkov, je *A* **sestavljen** iz teh dogodkov; če pa ga ni mogoče razčleniti na več nezdružljivih načinov, je dogodek *A* **elementaren** ali **izid**
- **obseg dogodkov** je množica dogodkov *M* z naslednjimi lastnostmi: če je v *M* dogodek *A*, je v *M* tudi $\overline{A}$;če sta v *M* dogodka *A* in *B*, je v *M* tudi njun produkt

3. Lastnosti verjetnosti

- $0 \le P(A) \le 1$
- $P(G) = 1$ in $P(N) = 0$
- $A \subset B \Rightarrow P(A) \le P(B)$
- $P(A + B) = P(A) + P(B) - P(AB)$
- $P(\overline{A}) = 1 - P(A)$
- če je od *s* mogočih in enako verjetnih izidov za dogodek *A* ugodnih *r* izidov, je $P(A) = \frac{r}{s}$; ta način določanja verjetnosti imenujemo **klasičen način**

4. Pogojna verjetnost

Pogojna verjetnost $P(A/B)$ ali $P_B(A)$ je verjetnost dogodka *A* pri pogoju, da se zgodi dogodek *B*. Verjetnost $P(A)$ pa je **brezpogojna** verjetnost dogodka *A*.
Za pogojno verjetnost velja naslednja formula:
$P(AB) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B)$ ali bolj splošno:
$P(A_1 A_2 \cdots A_n) = P(A_1) \cdot P(A_2 / A_1) \cdots P(A_n / A_1 A_2 \cdots A_{n-1})$
Dogodek *A* je **odvisen** od dogodka *B*, če je $P(A/B) \neq P(A)$ ter **neodvisen** od dogodka *B*, če je $P(A/B) = P(A)$. Če je dogodek *A* neodvisen od dogodka *B*, je tudi *B* neodvisen od *A*, pravimo, da sta dogodka neodvisna; takrat je $P(AB) = P(A) \cdot P(B)$ in $P(A_1 A_2 \cdots A_n) = P(A_1) \cdot P(A_2) \cdots P(A_n)$

5. Relejni poskusi

Relejni poskusi so poskusi, ki potekajo na več stopnjah in je poskus na naslednji stopnji odvisen od izidov na prejšnjih stopnjah. Najpreprostejši relejni poskus je poskus z dvema stopnjama. Pri takih poskusih velja **formula za popolno verjetnost** dogodka *A*: $P(A) = \sum_{i=1}^{n} [P(H_i) \cdot P(A/H_i)]$, kjer so dogodki $H_1, H_2, \ldots, H_i$ mogoči izidi na prvi stopnji, dogodek *A* pa je eden izmed mogočih dogodkov na drugi stopnji.

Verjetnost izida na prvi stopnji pod pogojem, da se je na drugi stopnji zgodil dogodek *A* pa se izračuna po naslednji formuli: $P(H_k / A) = \frac{P(H_k) \cdot P(A/H_k)}{P(A)}$

Če v tej formuli $P(A)$ zamenjamo z vsoto $\sum_{i=1}^{n} [P(H_i) \cdot P(A/H_i)]$ po formuli za popolno verjetnost, pa dobimo **Bayesovo formulo** ali **izrek o verjetnosti hipotez**:

$$P(H_k / A) = \frac{P(H_k) \cdot P(A/H_k)}{\sum_{i=1}^{n} [P(H_i) \cdot P(A/H_i)]}$$

6. Zaporedja neodvisnih poskusov

Več poskusov je med seboj neodvisnih, če naberemo iz vsakega od poskusov po en dogodek na kakršen koli način in so ti dogodki med seboj neodvisni. **Bernoullijevo zaporedje** je zaporedje enakih neodvisnih poskusov, v katerem se lahko zgodita v istem poskusu le dva dogodka, in sicer *A* z verjetnostjo *p* ($0 < p < 1$) ali njegova negacija $\overline{A}$ z verjetnostjo $q = 1 - p$. Verjetnost, da se bo dogodek *A* v *n* ponovitvah poskusa zgodil natanko *k*-krat izračunamo po **Bernoullijevi formuli**:
$P_n(k) = \binom{n}{k} \cdot p^k \cdot q^{n-k}$ pri tem pa velja: $\sum_{k=0}^{n} P_n(k) = 1$
Funkcija $P_n(k)$ ima največjo vrednost pri $k = np - q$ (oziroma prvem večjem celem številu, če to število ni celo) ter $k = np + p$. Dokler je *k* manjši od najverjetnejše vrednosti, funkcija strogo narašča, ko pa je večji, funkcija strogo pada.
Ker za zelo velike *n* Bernoullijeva formula ni uporabna, v teh primerih uporabljamo njen približek, **Laplaceovo lokalno formulo**: $P_n(k) \approx \frac{1}{\sqrt{2\pi npq}} \cdot e^{-\frac{(k-np)^2}{2npq}}$

Pripravnejšo obliko te funkcije dobimo, če vpeljemo sodo funkcijo $\varphi(x)$, kjer je: $\frac{k-np}{\sqrt{npq}}=x$ in $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$, ter tako dobimo formulo $P_n(k)\approx\frac{1}{\sqrt{npq}}\cdot\varphi(x)$

Vrednosti funkcije $\varphi(x)$ pa so zaradi njene pogoste uporabe tabelirane.

Ker je pri zelo majhnih ali zelo velikih *p* Laplaceova formula uporabna le pri precej velikih *n*, uporabljamo za razmeroma majhen *n* in majhen *p* **Poissonovo formulo**: $P_n(k)\approx\frac{(np)^k e^{-np}}{k!}$, ki jo lahko prilagodimo tudi za zelo velik *p*: $P_n(k)\approx\frac{(nq)^{n-k}e^{-nq}}{(n-k)!}$

Za računanje verjetnosti, da je $a\le k<b$, kjer je $0\le a<b\le n$, si pomagamo z **Laplaceovim integralskim izrekom**: Če je *p* verjetnost dogodka *A* in *k* njegova frekvenca v *n* ponovitvah poskusa, velja za poljubni realni števili α in β, kjer je $\alpha<\beta$, relacija: $\lim_{n\to\infty}P\left(\alpha\le\frac{k-np}{\sqrt{npq}}<\beta\right)=\frac{1}{\sqrt{2\pi}}\int_{\alpha}^{\beta}e^{-\frac{1}{2}x^2}dx$.

Če vpeljemo še liho funkcijo $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{1}{2}t^2}dt$, katere vrednosti so prav tako tabelirane, dobimo naslednjo formulo za računanje verjetnosti, da je $a\le k<b$: $P_n(a,b)\approx\Phi\left(\frac{b-np}{\sqrt{npq}}\right)-\Phi\left(\frac{a-np}{\sqrt{npq}}\right)$; formula je uporabna pri velikih *n*, pri majhnih *n* pa le, če se *p* malo loči od 0,5.

Pascalov obratni problem je izračunati, kolikšna je verjetnost, da se bo zgodil dogodek *A* v *n*-tem poskusu ravno *m*-tič, kjer je $n\ge m$. Izračuna se ga po naslednji formuli: $p_n(m)=\binom{n-1}{m-1}\cdot p^m\cdot q^{n-m}$

Če želimo pri poskusu s popolnim sistemom izidov ugotoviti, kolikšna je verjetnost $P_n(k_1,k_2,\ldots,k_r)$ dogodka, da se zgodi v *n* ponovitvah tega poskusa izid E_j natanko k_j -krat (*j* = 1, 2,..., *r*), kjer so $k_1,k_2,\ldots,k_r$ poljubna nenegativna cela števila in je njihova vsota enaka *n*, lahko uporabimo naslednjo formulo: $P_n(k_1,k_2,\ldots,k_r)=\frac{n!}{k_1!\cdot k_2!\cdots k_r!}\cdot p_1^{k_1}\cdot p_2^{k_2}\cdots p_r^{k_r}$

Ker pa je pri velikem *n* verjetnosti težko določiti po tej formuli, jih lahko aproksimiramo po formuli: $P_n(k_1,k_2,\ldots,k_r)=\frac{1}{\sqrt{(2\pi n)^{r-1}\cdot p_1\cdot p_2\cdots p_r}}\cdot e^{-\frac{2}{3}\left(q_1x_1^2+q_2x_2^2+\ldots+q_rx_r^2\right)}$, kjer velja za vsak *j* od 1 do *r*: $q_j=1-p_j$ in $x_j=\frac{k_j-np_j}{\sqrt{np_jq_j}}$

7. Slučajne spremenljivke

Slučajna spremenljivka je količina, katere vrednost je odvisna od slučaja; določena je s svojo **zalogo vrednosti** in s svojim **porazdelitvenim zakonom**.

Dogodek $(X=x)$ je pojav, da ima slučajna spremenljivka *X* vrednost *x*. Vsi mogoči dogodki $(X=x)$ sestavljajo popoln sistem.

Ločimo **diskretne** (zaloga vrednosti je zaporedje) in **zvezno porazdeljene** (zaloga vrednosti je interval) slučajne spremenljivke.

Porazdelitvena funkcija $F(x)$ slučajne spremenljivke *X* je funkcija, ki ima pri vsakem realnem *x* vrednost enako verjetnosti dogodka $(X < x)$: $F(x) = P(X < x)$
Lastnosti porazdelitvenih funkcij:

- vsaka porazdelitvena funkcija $F(x)$ je naraščajoča in zanjo veljata relaciji $F(-\infty) = \lim_{x\to-\infty} F(x) = 0$ ter $F(\infty) = \lim_{x\to\infty} F(x) = 1$
- za poljubni realni števili x_1 in x_2, kjer je $x_1 < x_2$ velja relacija $P(x_1 \le X < x_2) = F(x_2) - F(x_1)$
- vsaka porazdelitvena funkcija je od leve zvezna
- če je $F(x)$ porazdelitvena funkcija slučajne spremenljivke *X*, velja za vsak realen *x*: $F(x+0) - F(x) = P(X = x)$

Pri diskretnih slučajnih spremenljivkah je primernejša verjetnostna funkcija, ki je definirana takole: **Verjetnostna funkcija** p_k diskretne slučajne spremenljivke *X* ima pri vsakem mogočem *k* vrednost enako verjetnosti dogodka $(X = x_k)$: $p_k = P(X = x_k)$. Zapišemo jo lahko v t.i. **verjetnostni shemi**: $X : \begin{pmatrix} x_1 & x_2 & x_3 & \cdots \\ p_1 & p_2 & p_3 & \cdots \end{pmatrix}$

8. Diskretne porazdelitve

- **enakomerna diskretna porazdelitev**: slučajna spremenljivka zavzame vse vrednosti z enako verjetnostjo; $P(X = x_k) = \frac{1}{n}$ velja za vsak *k* od 1 do *n*
- **binomska porazdelitev**: slučajna spremenljivka *X* z zalogo vrednosti $\{0,1,\ldots,n\}$ ima verjetnostno funkcijo $p_k = \binom{n}{k} p^k (1-p)^{n-k}$, označimo jo z $b(n, p)$
- **Poissonova porazdelitev**: zaloga vrednosti je $\{0,1,2,\ldots\}$, verjetnostna funkcija pa je $p_k = \frac{a^k e^{-a}}{k!}$ $(a > 0)$, označimo jo s $P(a)$
- **Pascalova porazdelitev**: zaloga vrednosti je $\{m, m+1, m+2, \ldots\}$, verjetnostna funkcija pa je $p_k = \binom{k-1}{m-1} p^m (1-p)^{k-m}$; v primeru, da je $m = 1$, dobimo **geometrijsko porazdelitev**: $p_k = p(1-p)^{k-1}$ $(k = 1,2,3,\ldots)$

9. Zvezne porazdelitve

Gostota verjetnosti $p(x)$ je funkcija, ki jo dobimo z odvajanjem porazdelitvene funkcije: $p(x) = F'(x) \ge 0$. Za gostoto verjetnosti velja:

$$\int_{-\infty}^{\infty} p(x)dx = 1 \qquad P(x_1 \le X < x_2) = \int_{x_1}^{x_2} p(x)dx \qquad P(X = x) = \int_x^x p(t)dt = 0$$

- **enakomerna zvezna porazdelitev** ima gostoto $p(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{drugod} \end{cases}$; torej velja relacija $P(u < X < v) = \frac{v-u}{b-a}$, kjer je $a \le u < v \le b$

- **normalna ali Gaussova porazdelitev** ima gostoto $p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}$; označimo jo z $N(a,\sigma)$; krivulja $y=p(x)$ ima v točki $x=a$ teme, v točkah $x=a-\sigma$ in $x=a+\sigma$ pa prevoja. Verjetnost na nekem intervalu izračunamo takole: $P(x_1<X<x_2)=\Phi\left(\frac{x_2-a}{\sigma}\right)-\Phi\left(\frac{x_1-a}{\sigma}\right)$

Gostota najpreprostejše oz. **standardizirane normalne porazdelitve** $N(0,1)$ je $p(x)=\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$, poenostavi se tudi formula za verjetnost: $P(x_1<X<x_2)=\Phi(x_2)-\Phi(x_1)$

Za velike *n* je mogoče binomsko porazdelitev $b(n,p)$ aproksimirati z normalno porazdelitvijo $N(np,\sqrt{npq})$

- **porazdelitev** $\chi^2(n)$ ima gostoto $p(x)=\begin{cases}\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}} & x>0\\ 0 & x\le 0\end{cases}$, kjer je *n* poljubno naravno število oz. **število prostorskih stopenj**, izraz $\Gamma(x)$ pa je **Eulerjev integral** $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$

Če je *X* standardizirano normalno porazdeljena, je X^2 porazdeljena $\chi^2(1)$, po zakonu $\chi^2(n)$ pa je porazdeljena vsota kvadratov *n* standardizirano normalnih slučajnih spremenljivk, če so te med seboj neodvisne

- **Studentova porazdelitav** $S(n)$ ima gostoto $p(x)=\frac{1}{\sqrt{n}\cdot \mathrm{B}\left(\frac{n}{2},\frac{1}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$, kjer je *n* poljubno naravno število oz. **število prostorskih stopenj**, izraz $\mathrm{B}(x,y)$ pa je **Eulerjeva funkcija beta** $\mathrm{B}(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt$

Če sta spremenljivki *X* in *Y* neodvisni ter je *X* porazdeljena po $N(0,1)$ in *Y* porazdeljena po $\chi^2(n)$, je spremenljivka $Z=\frac{X\sqrt{n}}{\sqrt{Y}}$ porazdeljena po $S(n)$

- **Snedecorjeva porazdelitev** $F(m,n)$ ima gostoto $p(x)=\begin{cases}\frac{\sqrt{m^m n^n}}{\mathrm{B}\left(\frac{m}{2},\frac{n}{2}\right)}\sqrt{\frac{x^{m-2}}{(n+mx)^{m+n}}} & x>0\\ 0 & x\le 0\end{cases}$, kjer sta *m* in *n* poljubni naravni števili oz. **števili prostorskih stopenj**

Če sta spremenljivki *X* in *Y* neodvisni ter je *X* porazdeljena po $\chi^2(m)$ in *Y* porazdeljena po $\chi^2(n)$, je spremenljivka $Z=\frac{nX}{mY}$ porazdeljena po $F(m,n)$, spremenljivka Z^{-1} pa po $F(n,m)$; kvadrat spremenljivke, ki je porazdeljena po $S(n)$ pa je porazdeljen po $F(1,n)$

10. Osnove slučajnih vektorjev

Slučajen vektor dimenzije *n* je *n*-terica slučajnih spremenljivk $X = (X_1, X_2, \dots, X_n)$. Njegova zaloga vrednosti je neka množica v *n*-razsežnem prostoru $\boldsymbol{R}_n$, njegova porazdelitvena funkcija pa $F(x_1, x_2, \dots, x_n) = P(X_1 < x_1, X_2 < x_2, \dots, X_n < x_n)$. Pri slučajnem vektorju veljajo enake zakonitosti kot pri slučajni spremenljivki z nekaj novimi:

- $F(x_1, x_2, \dots, x_{n-1}, \infty) = P(X_1 < x_1, X_2 < x_2, \dots, X_{n-1} < x_{n-1})$
- $F(x_1, \infty, \dots, \infty) = P(X_1 < x_1) = F_1(x_1)$; ta relacija velja za vsako komponento vektorja *X*
- Naj bosta $a = \{a_1, a_2, \dots a_n\}$ in $b = \{b_1, b_2, \dots b_n\}$ taki točki v $\boldsymbol{R}_n$, da velja $\forall i \in \{1..n\} : a_i < b_i$, kar lahko zapišemo tudi kot $a < b$; potem velja $P(a \le X < b) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)$

Diskretno porazdeljen slučajni vektor lahko opišemo tudi z verjetnostno funkcijo $p(k_1, k_2, \dots, k_n) = P(X_1 = k_1, X_2 = k_2, \dots, X_n = k_n)$, kjer preteče vektor $(x_{k_1}, x_{k2}, \dots, x_{k_n})$ vso zalogo vrednosti slučajnega vektorja. **Zvezno porazdeljen** slučajni vektor ima v splošnem gostoto verjetnosti $p(x_1, x_2, \dots, x_n) = \dfrac{\partial^n F(x_1, x_2, \dots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}$.

***n*-razsežna normalna porazdelitev** ima gostoto verjetnosti $p(x) = \sqrt{\dfrac{|A|}{(2\pi)^n}} e^{-\frac{1}{2}(x-a)^T A (x-a)}$, kjer je $a = \begin{Vmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{Vmatrix}$ dan konstanten matrični stolpec, $x = \begin{Vmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Vmatrix}$ spremenljiv neslučajen matrični stolpec in $A = \begin{Vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{Vmatrix}$ konstantna simetrična pozitivno definitna matrika, $x^T = \|x_1, x_2, \dots, x_n\|$ pa transponirana matrika *x*.

11. Matematično upanje

Matematično upanje $E(X)$ je **povprečna vrednost** slučajne spremenljivke oz. vrednost, pri kateri se navadno stabilizira povprečje vrednosti slučajne spremenljivke v velikem številu realizacij. Je ena od **številskih karakteristik** slučajne spremenljivke. Če ima slučajna spremenljivka *X* porazdelitveno funkcijo $F(X)$, je $E(X) = \int_{-\infty}^{\infty} x\,dF(x)$.

- pri diskretni slučajni spremenljivki $X : \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$ je $E(X) = \sum_{i=1}^{n} x_i p_i$

  če je diskretna slučajna spremenljivka neomejena $(n = \infty)$, ima matematično upanje natanko takrat, kadar je $\sum_{i=1}^{\infty} |x_i| p_i < \infty$, sicer matematično upanje ne obstaja

- pri zvezno porazdeljeni slučajni spremenljivki pa je: $E(X)=\int_{-\infty}^{\infty} x\,p(x)dx$;

  matematično upanje obstaja takrat, kadar je $\int_{-\infty}^{\infty}|x|\,p(x)dx<\infty$

Za posamezne porazdelitve velja:

- Poissonova porazdelitev: $E(X)=a$
- binomska porazdelitev: $E(X)=np$
- enakomerna porazdelitev na intervalu $[a,b]$: $E(X)=\frac{a+b}{2}$
- normalna porazdelitev $N(a,\sigma)$: $E(X)=a$

12. Disperzija

Disperzija ali **varianca** slučajne spremenljivke *X*, $D(X)$, je razpršenost te slučajne spremenljivke okrog njenega matematičnega upanja. Za disperzijo velja za vse porazdelitve naslednja formula: $D(X)=E\left((X-E(X))^2\right)=E(X^2)-E(X)^2$. Torej velja za:

- diskretno slučajno spremenljivko $D(X)=\sum_i (x_i-E(X))^2 p_i$
- zvezno porazdeljeno slučajno spremenljivko $D(X)=\int_{-\infty}^{\infty}(x-E(X))^2 p(x)dx$

Splošno velja torej $D(X)=\int_{-\infty}^{\infty}(x-E(X))^2 dF(x)$, kjer je $F(X)$ porazdelitvena funkcija slučajne spremenljivke *X*.

Ker disperzijo računamo iz <u>kvadriranih</u> odklonov od povprečja, je ne moremo upoštevati kot mero razpršenosti. Zato za mero razpršenosti vzamemo pozitivni kvadratni koren iz disperzije oz. **standardno deviacijo** slučajne spremenljivke *X*: $\sigma(X)=\sigma_X=\sqrt{D(X)}$

Za posamezne porazdelitve velja:

- Poissonova porazdelitev: $E(X^2)=a^2+a$, torej je $D(X)=a^2+a-a^2=a$
- binomska porazdelitev: $D(X)=npq$
- enakomerna porazdelitev na intervalu $[a,b]$: $E(X^2)=\frac{a^2+ab+b^2}{3}$, torej je $D(X)=\frac{b^2+ab+a^2}{3}-\frac{b^2+2ab+a^2}{4}=\frac{(b-a)^2}{12}$
- normalna porazdelitev $N(a,\sigma)$: $D(X)=\sigma^2$

13. Lastnosti matematičnega upanja in disperzije

- če je *a* konstanta in je $P(X=a)=1$, je $E(X)=a$ in $D(X)=0$
- matematično upanje je linearen funkcional: če obstajata $E(X)$ in $E(Y)$ ter če sta *a* in *b* poljubni naravni števili, velja formula $E(aX+bY)=aE(X)+bE(Y)$;

  to formulo se lahko tudi posploši: $E\left(\sum_{k=1}^{n} a_k X_k\right)=\sum_{k=1}^{n} a_k E(X_k)$
- za vsako slučajno spremenljivko, ki ima mat. upanje, je $E(X-E(X))=0$
- če obstajata $E(X^2)$ in $E(Y^2)$, potem velja $|E(XY)|\le\sqrt{E(X^2)E(Y^2)}$
- če slučajni spremenljivki *X* in *Y* imata matematično upanje ter velja relacija $E(XY)=E(X)E(Y)$, sta *X* in *Y* med seboj **nekorelirani**, če velja relacija

$E(XY) \neq E(X)E(Y)$ pa sta **korelirani**; neodvisnost spremenljivk X in Y je zadosten pogoj za njuno koreliranost, ni pa potreben

- če obstaja $D(X)$ in je a poljubno realno število, velja $D(aX) = a^2 D(X)$ in $E\left((X-a)^2\right) \geq D(X)$
- **kovarianca** med slučajnima spremenljivkama X in Y $K(X,Y) = E\left((X - E(X))(Y - E(Y))\right) = E(XY) - E(X)E(Y)$ je matematično upanje produkta odklonov spremenljivk X in Y od njunih povprečnih vrednosti; kovarianca gotovo obstaja, če imata X in Y disperzijo; velja ocena $|K(X,Y)| \leq \sqrt{D(X)D(Y)} = \sigma(X) \cdot \sigma(Y)$
- slučajni spremenljivki X in Y sta nekorelirani natanko takrat, kadar je $K(X,Y) = 0$ [A]
- če imata slučajni spremenljivki X in Y disperzijo, velja naslednja formula: $D(X+Y) = D(X) + D(Y) + 2K(X,Y)$; to formulo se da tudi posplošiti: $D\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} D(X_k) + 2 \sum_{1 \leq i < j \leq n} K(X_i, X_j)$
- če imajo slučajne spremenljivke $X_1, X_2, \ldots, X_n$ disperzijo in so paroma nekorelirane, je $D\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} D(X_k)$, ker so vse kovariance enake 0

14. Korelacijski koeficient

Količina $|K(X,Y)|$ je tem večja, čim bolj sta X in Y linearno povezana ter čim bolj sta razpršeni. Če želimo odstraniti vpliv razpršenosti ter ugotoviti dejansko mero linearne odvisnosti med X in Y, uvedemo **korelacijski koeficient**: $r(X,Y) = \dfrac{K(X,Y)}{\sigma(X) \cdot \sigma(Y)}$.

Izrek z opombo iz prejšnjega poglavja torej lahko popravimo:
Slučajni spremenljivki X in Y sta nekorelirani natanko takrat, kadar je $r(X,Y) = 0$.
Za korelacijski koeficient velja: $-1 \leq r(X,Y) \leq 1$; relacija $|r(X,Y)| = 1$ pa velja natanko takrat, kadar je z verjetnostjo 1: $Y = aX + b$.

15. Osnovni pojmi matematične statistike

Osnovno vprašanje statistike je, kaj je mogoče reči o slučajni spremenljivki X na **populaciji** (končna ali neskončna množica vseh elementov, ki jih preučujemo), na podlagi informacije, ki smo jo dobili o njej iz **vzorca** (primerno izbran manjši del populacije). Najboljše izbran vzorec je **slučajni vzorec**, za katerega velja, da ima vsak element iz populacije enako možnost, da bo izbran.

16. Opisovanje statističnih podatkov

Naj bodo vrednosti slučajne spremenljivke X na slučajnem vzorcu velikosti n enake $x_1, x_2, \ldots, x_n$ ter vse mogoče različne vrednosti enake $x_1, x_2, \ldots, x_r$ s frekvencami k_i. Te podatke lahko zapišemo v tabeli, podobni kot je verjetnostna shema diskretne slučajne spremenljivke:

[A] glej naslednje poglavje za popolno definicijo

$$\begin{matrix} x_1 & x_2 & \cdots & x_r \\ k_1 & k_2 & \cdots & k_r \end{matrix}$$

Če je število *r* različnih vrednosti spremenljivke *X* preveliko, da bi bila zgornja tabela pregledna, grupiramo sosednje vrednosti v *m* **razredov** enake širine *d*. Ponavadi je *m* med 10 in 20, *d* pa je določen tako, da je dolžina intervala, na katerem so vse vrednosti enaka ali malo manjša od *md*. Spodnjo mejo razreda označimo z x_i', zgornjo z x_i'', ter sredino razreda z x_i. Spodnja meja še spada k razredu, zgornja pa ne. **Frekvenčno porazdelitev** lahko zapišemo enako kot prej (**v tabeli**), s **histogramom** ali s **frekvenčnim poligonom**.

Za take podatke je vzorčno povprečje $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{r} x_i k_i$ ter vzorčna disperzija $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{r}\left(x_i - \bar{x}\right)^2 k_i$; ker je tako računanje včasih zelo zamudno, vpeljemo **pomožno spremenljivko** $Y = X - c$, kjer *c* izberemo tako, da je približno na sredini med največjo in najmanjšo vrednostjo spremenljivke *X*. Potem je $\bar{x} = c + \bar{y}$ in $\sigma_x^2 = \sigma_y^2$.

17. Preskušanje statističnih hipotez

Statistična hipoteza je vsaka domneva o neznanem porazdelitvenem zakonu slučajne spremenljivke *X*. Hipoteze o vrednostih parametrov porazdelitve imenujemo **parametrične**, hipoteze o tipu porazdelitvenega zakona pa **neparametrične**. Če hipoteza natanko določa porazdelitveni zakon, se imenuje **enostavna**, sicer pa **sestavljena**. Vse hipoteze, ki pridejo v poštev pri danem problemu so **dopustne**; tisto, ki jo želimo preskusiti imenujemo **ničelna hipoteza** H_0, druge dopustne hipoteze pa so **alternativne**. **Preskus** ali **test** hipoteze H_0 je pravilo, po katerem hipotezo **sprejmemo** ali **zavrnemo**.

Zanima nas slučajna spremenljivka *X*, dana na populaciji *G*; edina informacija, ki jo imamo o *X* je slučajen vzorec $(x_1, x_2, \ldots, x_n)$, ki je realizacija slučajnega vektorja $(X_1, X_2, \ldots, X_n)$. Ta slučajni vektor ima neodvisne komponente, porazdeljene tako, kot je *X* na populaciji *G*. Naj bo *W* zaloga vrednosti tega slučajnega vektorja ali **prostor vzorcev**. Izberimo si neko množico $w \subset W$ in predpišimo: če bo $(X_1, X_2, \ldots, X_n) \in w$, bomo H_0 zavrnili, če bo $(X_1, X_2, \ldots, X_n) \in W - w$, pa bomo pustili vprašanje o pravilnosti H_0 odprto. Tak test imenujemo **preskus značilnosti**, množico *w* pa **kritično območje** testa. **Napako prve vrste** napravimo, če zavrnemo pravilno hipotezo H_0, **napako druge vrste** pa, če sprejmemo napačno hipotezo H_0. Ker v nobenem primeru ne sprejmemo hipoteze H_0, napake druge vrste praktično ne moremo narediti; verjetnost napake prve vrste pa je vnaprej predpisana vrednost α oz. **stopnja značilnosti** testa. Kadar H_0 zavrnemo, pravimo, da se H_0 **značilno loči** od eksperimentalnih podatkov oz. da je **razloček signifikanten**; kadar je ne zavrnemo, pa pravimo, da H_0 **ni v nasprotju z eksperimentalnimi podatki**.

Za konkreten primer si navadno izberemo primerno funkcijo komponent slučajnega vektorja $U = f(X_1, X_2, \ldots, X_n)$. Slučajno spremenljivko *U* imenujemo **statistika** vzorca $(X_1, X_2, \ldots, X_n)$. Nato si na realni osi izberemo tako množico w_u, da je $P(U \in w_u / H_0) \leq \alpha$ (verjetnost, da je $U \in w_u$ pri pogoju, da velja H_0 mora biti manjša ali enaka stopnji značiolnosti α). Potem je kritično območje *w* množica tistih točk iz prostora vzorcev *W*, ki jih ta funkcija upodobi v w_u : $w = f^{-1}(w_u)$. Množica w_u je najpogosteje interval.

18. Parametrični preskusi značilnosti

S parametričnimi preskusi značilnosti preskušamo **parametrične hipoteze**. Kritično območje w_u ima najpogosteje naslednje oblike:

- $w_u = \{u : |u| \geq u_\alpha\}$, kjer je u_α določen s pogojem $P(|U| \geq u_\alpha / H_0) \leq a$
- $w_u = \{u : u \geq u_\alpha\}$, kjer je u_α določen s pogojem $P(U \geq u_\alpha / H_0) \leq a$
- $w_u = \{u : u \leq u_\alpha\}$, kjer je u_α določen s pogojem $P(U \leq u_\alpha / H_0) \leq a$

Iz vzorca, ki imamo podanega izračunamo eksperimentalno vrednost testne statistike *U*: $u_e = f(x_1, x_2, \dots, x_n)$. Če je potem $u_e \in w_u$, hipotezo H_0 zavrnemo, sicer pa ne.

Da bi določili u_α, moramo poznati porazdelitveni zakon statistike *U*, in sicer pri majhnem vzorcu natanko, pri velikem pa se zadovoljimo z njeno limitno porazdelitvijo, ko rase $n \to \infty$. Točne porazdelitve statistik so znane le zaprimer, ko je *X* na populaciji normalno porazdeljena, limitne porazdelitve pa obstajajo vedno.

Primeri preskusov značilnosti:

- *X* je na populaciji normalno porazdeljena z disperzijo σ^2 , njeno matematično upanje *a* pa nam ni znano; preskusiti želimo $H_0(a = a_0)$ z vzorcem velikosti *n* na stopnji značilnosti α :

  statistiko *U* (označimo jo z $\overline{X}$) si izberemo takole: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$;

  ta statistika je porazdeljena po zakonu $N\left(a_0, \frac{\sigma}{\sqrt{n}}\right)$;

  za testno statistiko vzamemo standarizirano normalno slučajno spremenljivko *Z*, ki jo dobimo iz $\overline{X}$ takole: $Z = \frac{\overline{X} - a_0}{\sigma}\sqrt{n}$;

  kritično območje bo $\{z : |z| \geq z_\alpha\}$, kjer je z_α rešitev enačbe $P(|Z| \geq z_\alpha) = \alpha$

- *X* je na populaciji normalno porazdeljena, ne poznamo pa njenega matematičnega upanja *a* in disperzije σ ; preskusiti želimo $H_0(a = a_0)$ z vzorcem velikosti *n* na stopnji značilnosti α :

  izberemo si statistiki: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ in $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, kjer je S^2 **cenilka za disperzijo** σ^2 in je $E(S^2) = \sigma^2$;

  ta statistika je porazdeljena po zakonu $N\left(a_0, \frac{\sigma}{\sqrt{n}}\right)$;

  za testno statistiko vzamemo slučajno spremenljivko *T* porazdeljeno po Studentovem zakonu $S(n-1)$, ki jo dobimo iz $\overline{X}$ in S^2 takole: $T = \frac{\overline{X} - a_0}{S}\sqrt{n}$;

  kritično območje bo $\{t : |t| \geq t_\alpha\}$, kjer je t_α rešitev enačbe $P(|T| \geq t_\alpha) = \alpha$

- X je na populaciji normalno porazdeljena, ne poznamo njenega matematičnega upanja a in disperzije σ; preskusiti želimo $H_0(\sigma = \sigma_0)$ z vzorcem velikosti n na stopnji značilnosti α:

izberemo si testno statistiko: $\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$;
ta statistika je porazdeljena po zakonu $\chi^2(n-1)$;
če je alternativna hipoteza $H_1(\sigma \neq \sigma_0)$ (test je **dvostranski**), bo kritično območje $\left\{\chi^2 : \chi^2 \le \chi^2_{\alpha/2} \text{ ali } \chi^2 : \chi^2 \ge \chi^{*2}_{\alpha/2}\right\}$, kjer je $\chi^2_{\alpha/2}$ rešitev enačbe

$P\left(\chi^2 \le \chi^2_{\alpha/2}\right) = \frac{\alpha}{2}$ in $\chi^{*2}_{\alpha/2}$ rešitev enačbe $P\left(\chi^2 \ge \chi^{*2}_{\alpha/2}\right) = \frac{\alpha}{2}$;

če pa je alternativna hipoteza $H_1(\sigma > \sigma_0)$ (test je **enostranski**), bo kritično območje $\left\{\chi^2 : \chi^2 \ge \chi^2_\alpha\right\}$, kjer je χ^2_α rešitev enačbe $P\left(\chi^2 \ge \chi^2_\alpha\right) = \alpha$;

- dani sta dve populaciji; X je porazdeljena po zakonu $N(a_1, \sigma)$, Y pa po zakonu $N(a_2, \sigma)$; za X imamo slučajen vzorec velikosti m, za Y pa velikosti n; preskusiti želimo $H_0(a_1 = a_2)$ na stopnji značilnosti α:

izberemo si statistike: $\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ in $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ ter

$$S^2 = \frac{1}{m+n-1}\left(\sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2\right);$$

za testno statistiko vzamemo slučajno spremenljivko T, ki jo dobimo takole: $T = \frac{\overline{X} - \overline{Y}}{S}\sqrt{\frac{mn}{m+n}}$;
ta statistika je porazdeljena po Studentovem zakonu $S(m+n-2)$;
kritično območje bo $\{t : |t| \ge t_\alpha\}$, kjer je t_α rešitev enačbe $P(|T| \ge t_\alpha) = \alpha$;
za ta primer ni nujno, da sta disperziji spremenljivk X in Y popolnoma enaki, dovolj je že, če sta vsaj približno enaki

- dani sta dve normalno porazdeljeni populaciji X in Y; za X imamo slučajen vzorec velikosti m, za Y pa velikosti n; preskusiti želimo $H_0(\sigma_1 = \sigma_2)$ na stopnji značilnosti α:

izberemo si statistike: $\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ in $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ ter

$$S_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2 \text{ in } S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2;$$

za testno statistiko vzamemo spremenljivko F, ki jo dobimo takole: $F = \frac{S_x^2}{S_y^2}$;

ta statistika je porazdeljena po Snedecorjevem zakonu $F(m-1, n-1)$;
kritično območje bo $\{F : |F| \ge F_\alpha\}$, kjer je F_α rešitev enačbe $P(|F| \ge F_\alpha) = \alpha$;

za F_e vzamemo vedno tak kvocient med s_x^2 in s_y^2, da je $F_e \geq 1$

19. Literatura

Rajko Jamnik, Verjetnostni račun in statistika, Društvo matematikov, fizikov in astronomov RS, 1986