I.5. Slučajne spremenljivke in porazdelitve

Denimo, da imamo poskus, katerega izidi so števila (npr. pri metu kocke so izidi števila pik). Se pravi, da je poskusom prirejena neka količina, ki more imeti različne vrednosti. Torej je spremenljivka. Katero od mogočih vrednosti zavzame v določeni ponovitvi poskusa, je odvisno od slučaja. Zato ji rečemo **slučajna spremenljivka**.

Da je slučajna spremenljivka znana, je potrebno vedeti

1. kakšne vrednosti more imeti (*zaloga vrednosti*) in
2. kolikšna je verjetnost vsake izmed možnih vrednosti ali intervala vrednosti.
   Predpis, ki določa te verjetnosti, imenujemo **porazdelitveni zakon**.

. . . Slučajne spremenljivke

Slučajne spremenljivke označujemo z velikimi tiskanimi črkami iz konca abecede, vrednosti spremenljivke pa z enakimi malimi črkami. Tako je npr. $(X = x_i)$ dogodek, da slučajna spremenljivka X zavzame vrednost x_i.

Porazdelitveni zakon slučajne spremenljivke X je poznan, če je mogoče za vsako realno število x določiti verjetnost

$$F(x) = P(X < x).$$

$F(x)$ imenujemo **porazdelitvena funkcija**.

Najpogosteje uporabljamo naslednji vrsti slučajnih spremenljivk:

1. **diskretna** slučajna spremenljivka, pri kateri je zaloga vrednosti neka števna (diskretna) množica;
2. **zvezna** slučajna spremenljivka, ki lahko zavzame vsako realno število znotraj določenega intervala.

Lastnosti porazdelitvene funkcije

1. *Funkcija F je definirana na vsem $\mathbb{R}$ in velja $0 \le F(x) \le 1, x \in \mathbb{R}$.*
2. *Funkcija F je nepadajoča $x_1 < x_2 \Longrightarrow F(x_1) \le F(x_2)$.*
3. *$F(-\infty) = 0$ in $F(\infty) = 1$.*
4. *Funkcija je v vsaki točki zvezna od leve $F(x-) = F(x)$.*
5. *Funkcija ima lahko v nekaterih točkah skok.*
   Vseh skokov je največ števno mnogo.
6. *$P(x_1 \le X < x_2) = F(x_2) - F(x_1)$.*
7. *$P(x_1 < X < x_2) = F(x_2) - F(x_1+)$.*
8. *$P(X \ge x) = 1 - F(x)$.*
9. *$P(X = x) = F(x+) - F(x)$.*

Diskretne slučajne spremenljivke

Zaloga vrednosti diskretne slučajne spremenljivke X je števna množica $\{x_1, x_2, \ldots, x_m, \ldots\}$. Dogodki

$$X = x_k \quad k = 1, 2, \ldots$$

sestavljajo popoln sistem dogodkov. Označimo verjetnost posameznega dogodka s

$$P(X = x_i) = p_i.$$

Vsota verjetnosti vseh dogodkov je enaka 1:

$$p_1 + p_2 + \cdots + p_m + \cdots = 1.$$

Verjetnostna tabela

Verjetnostna tabela prikazuje diskretno slučajno spremenljivko s tabelo tako, da so v prvi vrstici zapisane vse vrednosti x_i, pod njimi pa so pripisane pripadajoče verjetnosti:

$$X : \begin{pmatrix} x_1 & x_2 & \cdots & x_m & \cdots \\ p_1 & p_2 & \cdots & p_m & \cdots \end{pmatrix}.$$

Porazdelitvena funkcija je v tem primeru

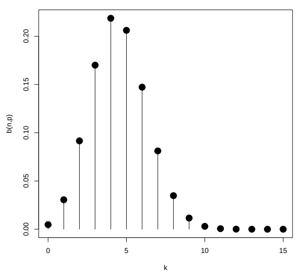
$$F(x_k) = P(X < x_k) = \sum_{i=1}^{k-1} p_i.$$

Enakomerna diskretna porazdelitev

Končna diskretna slučajna spremenljivka se porazdeljuje **enakomerno**, če so vse njene vrednosti enako verjetne. Primer take slučajne spremenljivke je število pik pri metu kocke

$$X: \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

Binomska porazdelitev




```
> h <- dbinom(0:15,size=15,prob=0.3)
> plot(0:15,h,type="h",xlab="k",ylab="b(n,p)")
> points(0:15,h,pch=16,cex=2)
```

Binomska porazdelitev ima zalogo vrednosti $\{0, 1, 2, \cdots, n\}$ in verjetnosti, ki jih računamo po Bernoullijevem obrazcu:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

$k = 0, 1, 2, \cdots, n$. Binomska porazdelitev je natanko določena z dvema podatkoma – parametroma: n in p. Če se slučajna spremenljivka X porazdeljuje binomsko s parametroma n in p, zapišemo:

$$X: \ B(n, p).$$

Binomska porazdelitev / Primer

Naj bo slučajna spremenljivka X določena s številom fantkov v družini s 4 otroki. Denimo, da je enako verjetno, da se v družini rodi fantek ali deklica:

$$P(F) = p = \frac{1}{2} \quad , \quad P(D) = q = \frac{1}{2}.$$

Spremenljivka X se tedaj porazdeljuje binomsko $B(4, \frac{1}{2})$ in njena verjetnostna shema je:

$$X: \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1/16 & 4/16 & 6/16 & 4/16 & 1/16 \end{pmatrix}.$$

Npr.

$$P(X = 2) = P_4(2) = \binom{4}{2}\left(\frac{1}{2}\right)^2 (1 - \frac{1}{2})^{4-2} = \frac{6}{16}$$

Porazdelitev obravnavane slučajne spremenljivke je simetrična.
Pokazati se da, da je binomska porazdelitev simetrična, le če je $p = 0,5$.

Poissonova porazdelitev $P(\lambda)$

Poissonova porazdelitev ima zalogo vrednosti $\{0, 1, 2, \ldots\}$, njena verjetnostna funkcija pa je

$$p_k = P(\#\texttt{dogodkov} = k) = \lambda^k \frac{e^{-\lambda}}{k!},$$

kjer je $\lambda > 0$ dani parameter
– pogostost nekega dogodka.

Posebno pomembna je v teoriji množične strežbe.

$$p_{k+1} = \frac{\lambda}{k+1} p_k, \quad p_0 = e^{-\lambda}.$$

Pascalova porazdelitev $P(m, p)$

Pascalova porazdelitev ima zalogo vrednosti $\{m, m+1, m+2, \ldots\}$, verjetnostna funkcija pa je

$$p_k = \binom{k-1}{m-1} p^m q^{k-m},$$

kjer je $0 < p < 1$ dani parameter
– verjetnost dogodka A v posameznem poskusu.

Opisuje porazdelitev števila poskusov potrebnih, da se dogodek A zgodi m–krat.

Za $m = 1$, porazdelitvi $G(p) = P(1, p)$ pravimo **geometrijska** porazdelitev. Opisuje porazdelitev števila poskusov potrebnih, da se dogodek A zgodi prvič.

Primer: Če mečemo kovanec toliko časa, da pade grb, in z X označimo število potrebnih metov, vključno z zadnjim, potem je slučajna spremenljivka X geometrijsko porazdeljena.

Če z X označimo število metov, vključno z zadnjim, do m-tega grba, potem dobimo **negativno binomsko** slučajno spremenljivko X : NegBin(m, p) in

$$P(X = k) = \binom{k-1}{m-1} p^m (1-p)^{k-m} \quad \text{za } k \geq m.$$

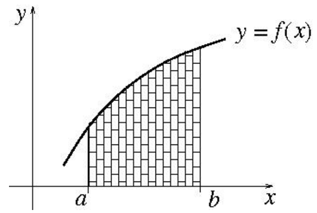
Hipergeometrijska porazdelitev $H(n; M, N)$

Hipergeometrijska porazdelitev ima zalogo vrednosti $\{0, 1, 2, \ldots\}$, verjetnostna funkcija pa je

$$p_k = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}},$$

kjer so $k \leq n \leq \min(M, N-M)$ dani parametri.
Opisuje verjetnost dogodka, da je med n izbranimi kroglicami natanko k belih, če je v posodi M belih in $N - M$ črnih kroglic in izbiramo n–krat brez vračanja.

Ponovitev: integrali



Določeni integral predstavlja ploščino pod krivuljo.

Naj bo funkcija $y = f(x)$ zvezna na $[a, b]$ in nenegativna.
Ploščina lika med krivuljo $f(x) \geq 0$, in abscisno osjo na intervalu $[a, b]$ je enaka določenemu integralu

$$\int_a^b f(x)dx.$$

Lastnosti določenega integrala:

1) $\displaystyle\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx.$

2) Če je $f(x) \leq 0 \;\; \forall x \in [a, b]$,

je vrednost integrala negativna.

3) Naj bo $c \in [a, b]$

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

4) Naj bo $f(x) \geq g(x),\;\; x \in [a, b]$,

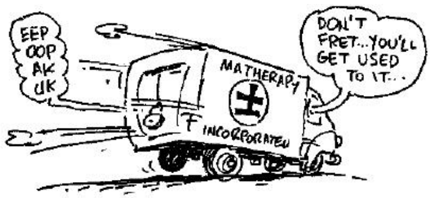
potem velja $\displaystyle\int_a^b f(x)\,dx \geq \int_a^b g(x)\,dx.$

Saj vas razumem!

Potem pa uporabimo še ∞ za mejo pri integriranju.

Brez preplaha!

Iščemo le celotno ploščino pod krivuljo,
od enega konca do drugega,
le da konca pravzaprav sploh ni.



Zvezne slučajne spremenljivke

Slučajna spremenljivka X je **zvezno porazdeljena**, če obstaja taka integrabilna funkcija p, imenovana **gostota verjetnosti**, da za vsak $x \in \mathbb{R}$ velja:

$$F(x) = P(X < x) = \int_{-\infty}^{x} p(t)\,dt,$$

kjer $p(x) \geq 0$. To verjetnost si lahko predstavimo tudi grafično v koordinatnem sistemu, kjer na abscisno os nanašamo vrednosti slučajne spremenljivke, na ordinatno pa gostoto verjetnosti $p(x)$. Verjetnost je tedaj predstavljena kot ploščina pod krivuljo, ki jo določa $p(x)$. *Velja*

$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \quad \text{in} \quad P(x_1 \leq X < x_2) = \int_{x_1}^{x_2} p(t)\,dt$$

ter $p(x) = F'(x)$.

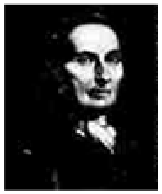
Enakomerna porazdelitev zvezne slučajne spremenljivke

Verjetnostna gostota enakomerno porazdeljene zvezne slučajne spremenljivke je:

$$p(x) = \begin{cases} \dfrac{1}{b-a} & a \leq X \leq b \\ 0 & \text{drugod.} \end{cases}$$

Grafično si jo predstavljamo kot pravokotnik nad intervalom (a, b) višine $\dfrac{1}{b-a}$.

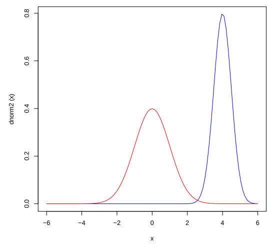
Normalna ali Gaussova porazdelitev



Leta 1738 je Abraham De Moivre (1667-1754) objavil aproksimacijo binomske porazdelitve, ki je normalna krivulja.

Leta 1809 je Karl Frederic Gauss (1777-1855) raziskoval matematično ozadje planetarnih orbit, ko je izpeljal normalno porazdelitveno funkcijo.

... Normalna porazdelitev




```
> d2 <- function(x){dnorm(x,mean=4,sd=0.5)}
> curve(d2,-6,6,col="blue")
> curve(dnorm,-6,6,col="red",add=TRUE)
```

Zaloga vrednosti **normalno porazdeljene** slučajne spremenljivke so vsa realna števila, gostota verjetnosti pa je:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

Normalna porazdelitev je natanko določena z parametroma: μ in σ.

Če se slučajna spremenljivka X porazdeljuje normalno s parametroma μ in σ zapišemo:

$$X : N(\mu, \sigma).$$

Laplaceov intervalski obrazec

Zanima nas, kolikšna je verjetnost $P_n(k_1, k_2)$, da se v Bernoullijevem zaporedju neodvisnih poskusov v n zaporednih poskusih zgodi dogodek A vsaj k_1–krat in manj kot k_2–krat. Označimo

$$x_k = \frac{k - np}{\sqrt{npq}} \quad \text{in} \quad \Delta x_k = x_{k+1} - x_k = \frac{1}{\sqrt{npq}}.$$

Tedaj je, če upoštevamo Laplaceov točkovni obrazec,

$$P_n(k_1, k_2) = \sum_{k=k_1}^{k_2-1} P_n(k) = \frac{1}{\sqrt{2\pi}} \sum_{k=k_1}^{k_2-1} e^{-\frac{1}{2}x_k^2} \Delta x_k.$$

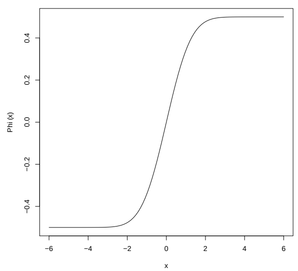
Za (zelo) velike n lahko vsoto zamenjamo z integralom

$$P_n(k_1, k_2) \approx \frac{1}{\sqrt{2\pi}} \int_{x_{k_1}}^{x_{k_2}} e^{-\frac{1}{2}x^2} dx.$$

Funkcija napake $\Phi(x)$

Funkcija napake imenujemo funkcijo

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}t^2} dt.$$




```
> Phi <- function(x){pnorm(x)-0.5}
> curve(Phi,-6.6)
```

Funkcija napake je liha, zvezno odvedljiva, strogo naraščajoča funkcija. $\Phi(-\infty) = -\frac{1}{2}$, $\Phi(0) = 0$, $\Phi(\infty) = \frac{1}{2}$ in $P_n(k_1, k_2) \approx \Phi(x_{k_2}) - \Phi(x_{k_1})$. Vrednosti funkcije napake najdemo v tabelah ali pa je vgrajena v statističnih programih.

```
> x2 <- (50 - 1000*0.05)/sqrt(1000*0.05*0.95)
> x1 <- (0 - 1000*0.05)/sqrt(1000*0.05*0.95)
> pnorm(x2)-pnorm(x1)
[1] 0.5
```


...Normalna porazdelitev

$$\begin{aligned} F(x) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{1}{2}s^2} ds \\ &= \frac{1}{2} + \Phi(\frac{x-\mu}{\sigma}). \end{aligned}$$

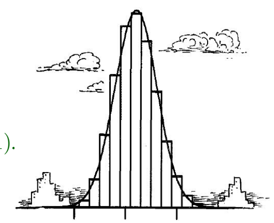
$$P(x_1 \le X < x_2) = \Phi\left(\frac{x_2 - \mu}{\sigma}\right) - \Phi\left(\frac{x_1 - \mu}{\sigma}\right)$$

Porazdelitev $N(0,1)$ je **standardizirana normalna porazdelitev**.

Spremenljivko $X : N(\mu, \sigma)$ *pretvorimo z*

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

v standardizirano spremenljivko $Z : N(0,1)$.



Iz Laplaceovega obrazca izhaja

$$B(n,p) \approx N(np, \sqrt{npq}).$$

Porazdelitev Poissonovega toka, eksponentna

Gostota **eksponentne porazdelitve** je enaka

$$p(x) = \lambda e^{-\lambda x}, \quad x \geq 0,$$

porazdelitvena funkcija pa

$$F(x) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}.$$

Porazdelitev gama

Naj bosta $b, c > 0$. Tedaj ima **porazdelitev Gama** $\Gamma(b,c)$ gostoto:

$$p(x) = \frac{c^b}{\Gamma(b)} x^{b-1} e^{-cx}, \ 0 < x$$

in $p(x) = 0$ za $x \leq 0$.

Funkcijo Gama lahko definiramo z določenim integralom za $\Re[z] > 0$ (Eulerjeva integralna forma)

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt = 2 \int_0^\infty e^{-t^2} t^{2z-1}\, dt,$$

Torej je $\Gamma(1) = 1$ in

$$\Gamma(z) = \int_0^1 \left[\ln \frac{1}{t}\right]^{z-1} dt.$$

(Je povsod analitična z izjemo $z = 0, -1, -2, \ldots$ in nima ničel.

Glej npr. `http://en.wikipedia.org/wiki/Gamma_function` in `http://en.wikipedia.org/wiki/Gamma_distribution`.)

Integracija po delih (po realnem argumentu) nam da (za $v = t^n$ in $du = e^{-t}dx$ velja $dv = nt^{n-1}$ in $u = e^{-t}$):

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$$

$$= \left[-t^{x-1}e^{-t}\right]_0^\infty + \int_0^\infty (x-1) t^{x-2} e^{-t}\, dt$$

$$= (x-1) \int_0^\infty t^{x-2} e^{-t}\, dt = (x-1)\Gamma(x-1).$$

Za naravno število x ($n = 1, 2, 3, \ldots$), dobimo

$$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) = (n-1)(n-2)\ldots 1 = (n-1)!$$

torej se Γ funkcija zreducira v 'faktorijel'.

Porazdelitev hi-kvadrat

Porazdelitev **hi-kvadrat** je poseben primer porazdelitve Gama:

$$\chi^2(n) = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

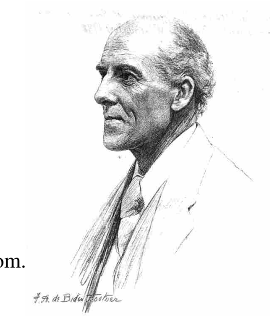
($n \in \mathbb{N}$ je število prostostnih stopenj) in ima gostoto

$$p(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \quad \text{kjer je} x > 0.$$

Ernst Abbe (1840-1905)

Leta 1863 jo je prvi izpeljal nemški fizik Ernst Abbe, ko je preučeval porazdelitev vsote kvadratov napak.

Leta 1878 je Ludwig Boltzmann izpeljal hi-kvadrat porazdelitev z dvema in tremi prostostnimi stopnjami, ko je študiral kinetično energijo molekul.



Karl Pearson (1875-1937) je demonstriral uporabnost hi-kvadrat porazdelitve statistikom.

Cauchyeva porazdelitev

z gostoto

$$p(x) = \frac{a}{\pi} \frac{1}{1 + a^2(x-b)^2}, \quad -\infty < x < \infty,\ a > 0$$

ima porazdelitveno funkcijo

$$F(x) = \frac{a}{\pi} \int_{-\infty}^{x} \frac{1}{1 + a^2(x-b)^2} \mathrm{d}x = \frac{1}{\pi} \operatorname{arc\,tg}(a(x-b)) + \frac{1}{2}$$

Porazdelitve v R-ju

V R-ju so za delo s pomembnejšimi porazdelitvami na voljo funkcije:

`d`*ime* – gostota porazdelitve *ime* $p_{ime}(x)$

`p`*ime* – porazdelitvena funkcija *ime* $F_{ime}(q)$

`q`*ime* – obratna funkcija: $q = F_{ime}(p)$

`r`*ime* – slučajno zaporedje iz dane porazdelitve

Za *ime* lahko postavimo: `unif` – zvezna enakomerna, `binom` – binomska, `norm` – normalna, `exp` – eksponentna, `lnorm` – logaritmičnonormalna, `chisq` – porazdelitev χ^2, ...

Opis posamezne funkcije in njenih parametrov dobimo z ukazom `help`. Na primer `help(rnorm)`.