VAJE 3: Generiranje statističnih spremenljivk

Na računalniških vajah se za urejanje in prikazovanje statističnih podatkov uporabi statistični programski paket SPSS in podatkovna datoteka *podatki3.sav*.

NALOGE:

1. Ustvari statistično spremenljivko *RStarost*. Osebam starim med 10 in 19 let določi vrednost 1, med 20 in 29 let vrednost 2, med 30 in 39 let vrednost 3 in med 40 in 49 let vrednost 4 (uporabi postopek *Transform - Compute Variables*, v *Target Variable & Label* uredi spremenljivko, nato uporabi *If.. (optional case selection condition)*). Spremenljivki tudi ustrezno popravi stolpec *Values* v zavihku *Variable View*.

2. Ustvari statistično spremenljivko *Vsota*, ki sešteje tisočice, stotice, desetice in enice statistične spremenljivke *KolicinaTD* (uporabi postopek *Transform - Compute Variables*, v *Target Variable & Label* uredi spremenljivko, v *Function group* izberi *Arithmetic* ter v stolpcu *Functions and Special Variables* za rešitev naloge uporabi funkcijo *Mod*). Opomba: lahko si pomagaš z dodatnimi spremenljivkami, npr. *Tisocice*, *Stotice*, *Desetice* in *Enice*.

3. Ustvari statistično spremenljivko *ITM* (indeks telesne mase) za vse osebe v vzorcu (formula je $k = \frac{m}{v^2}$, kjer je m masa v kilogramih in v višina v metrih). Ugotovi, koliko oseb je normalno prehranjenih (ima k med $18,5$ in $24,9$). Uporabi postopek *Transform - Compute Variables*, v *Target Variable & Label* uredi spremenljivko in za izračun uporabi računalo. Pomoč: število normalno prehranjenih oseb v vzorcu lahko ugotoviš s postopkom *Data - Select Cases*.

4. Privzemimo, da je starost oseb v vzorcu porazdeljena normalno. Ustvari statistično spremenljivko, ki pove, s kolikšno verjetnostjo se pojavijo posamezne starosti iz vzorca. Določi tudi komulativno verjetnost za statistično spremenljivko *Starost* (uporabi postopek *Transform - Compute Variables*, v *Target Variable & Label* uredi spremenljivko, v *Functon group* za izračun verjetnosti izberi *PDF & Noncentral PDF* ter v stolpcu *Functions and Special Variables* uporabi funkcijo *PDF.Normal*, za izračun komulativne verjetnosti pa *CDF & Noncentral CDF* ter v stolpcu *Functions and Special Variables* uporabi *CDF.Normal*). Opomba: najprej izračunaj vzorčno povprečje in vzorčni standardni odklon statistične spremenljivke *Starost* in dobljena podatka uporabi v zgornjem postopku.

5. Izračunaj vzorčno povprečje in vzorčni standardni odklon spremenljivke *TezaO*. Na podlagi dobljenih podatkov ustvari novo statistično spremenljivko *GTezaO*, ki naj bo porazdeljena normalno. Dobljene rezultate zaokroži na najbližje celo število (uporabi postopek *Transform - Compute Variables*, v *Target Variable & Label* uredi spremenljivko, v *Function group* izberi *Random Numbers* ter v *Functions and Special Variable* uporabi generator *Rv.Normal*). Opomba: za zaokrožanje v *Function group* izberi *Arithmetic* ter v *Functions and Special Variable* uporabi funkcijo *Rnd*. Za spremenljivki $GTezaO$ in $VisinaO$ ponovi 3. nalogo.

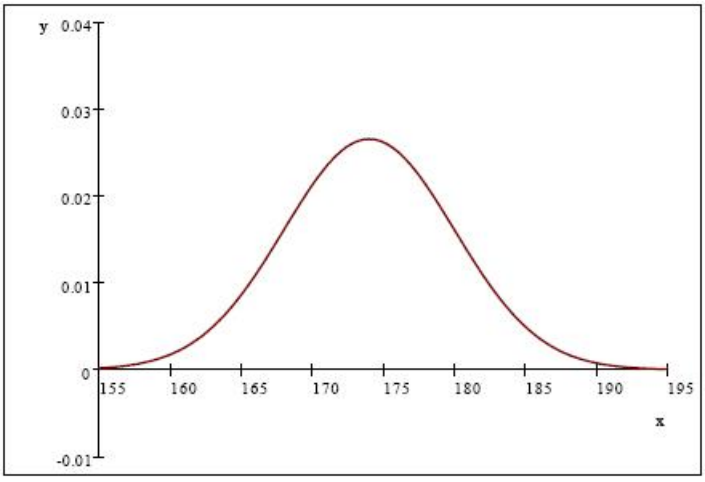
Teoretično ozadje

Normalna ali Gaussova porazdelitev

Najpomembnejša zvezna porazdelitev je t.i. normalna ali Gaussova porazdelitev $N(\mu, \sigma)$, ki je definirana z gostoto

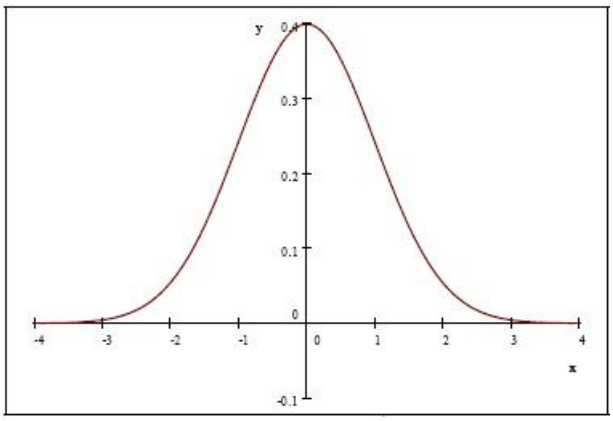
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

kjer je parameter μ matematično upanje in σ standardni odklon. Npr. telesna višina pri odraslih moških je na populaciji porazdeljena normalno s povprečjem $\mu = 172$ cm in standardnim odklonom $\sigma = 6$ cm. Glej sliko Telesna višina:



Slika 1: Telesna višina

Osnovno vprašanje, ki nas posebej zanima je denimo, kolikšen delež populacije ima telesno višino med 166 in 178 centimetri, kar predstavlja osrednji del populacije, ki je od povprečja oddaljen za en standardni odklon. V ta namen vpeljemo t.i. standardizirano normalno porazdelitev. Standardizirana normalna porazdelitev je najpreprostejša normalna porazdelitev $N(0,1)$ z matematičnim upanjem 0 in standardnim odklonom 1.



Slika 2: $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

Če je slučajna spremenljivka X porazdeljena normalno $N(\mu, \sigma)$, potem je slučajna spremenljivka

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

porazdeljena standardizirano normalno. Zato za računanje verjetnosti pri normalnih porazdelitvah zadostuje poznavanje samo standardizirane normalne porazdelitve. Pri standardizirani normalni porazdelitvi Z je verjetnost, da spremenljivka zavzame vrednost z intervala $[0, z]$, kjer je $z \geq 0$, enaka

$$\Phi(z) = \int_0^z \varphi(x)\, dx$$

Vrednosti funkcije $\Phi(z)$ so tabelirane v prilogi *Tabela A*. Funkcijo Φ razširimo še na negativna števila s predpisom

$$\Phi(-z) = -\Phi(z)\,.$$

Potem je verjetnost, da Z zavzame vrednosti z intervala $[a, b]$ enaka

$$P[a \leq Z \leq b] = \Phi(b) - \Phi(a)\,.$$

Npr. kolikšen delež populacije ima telesno višino, ki je porazdeljena normalno $X \sim N(172, 6)$, med 166 in 178 centimetri? Po zgornji formuli je slučajna spremenljivka

$$Z = \frac{X - 172}{6}$$

porazdeljena standardizirno normalno. Zato lahko zapišemo

$$\begin{aligned} P[166 \leq X \leq 178] &= P\left[\frac{166 - 172}{6} \leq \frac{X - 172}{6} \leq \frac{178 - 172}{6}\right] \\ &= P[-1 \leq Z \leq 1] = \Phi(1) - \Phi(-1) \\ &= 2\Phi(1) = 0.6826, \end{aligned}$$

kjer iz tabele A preberemo $\Phi(1) = 0.3413$. Zato je ta delež približno 68%.

Druge pomembne porazdelitve

V nadaljevanju se bomo poleg normalne porazdelitve pri statističnem ocenjevanju parametrov in statističnem sklepanju (preiskušanju statističnih hipotez) srečali še z naslednjimi zveznimi porazdelitvami:

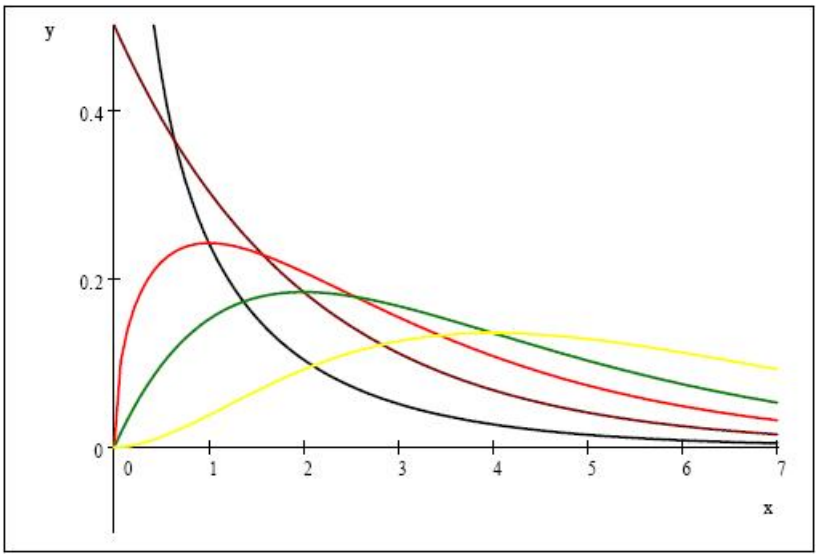
Porazdelitev hi kvadrat, ki jo označujemo z $\chi^2(n)$, ima gostoto verjetnosti [2, stran 42, zaznamek (10)]. Ta gostota je odvisna od parametra n, ki je poljubno naravno število. Rečemu mu število prostostnih stopenj.

Izkaže se, da če je spremenljivka X porazdeljena po zakonu $\chi^2(n)$, potem je spremenljivka $Y = \sqrt{2X} - \sqrt{2n - 1}$ pri velikih n porazdeljena približno standardizirano normalno.

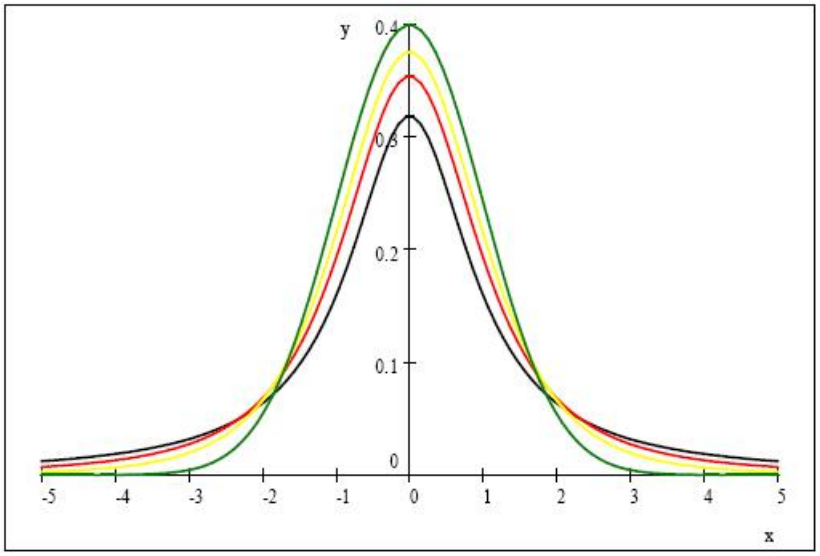
Studentova porazdelitev, ki jo označujemo s $S(n)$, ima gostoto verjetnosti [2, stran 43, zaznamek (11)]. Gostota je spet odvisna od naravnega števila n, ki mu rečemo število prostostnih stopenj.

Izkaže se, da lahko pri velikih n Studentovo porazdelitev $S(n)$ aproksimiramo s standardizirano normalno porazdelitvijo $N(0, 1)$.

Snedecorjeva porazdelitev, ki jo označujemo z $F(m, n)$, ima gostoto verjetnosti [2, stran 43, zaznamek (12)]. Gostota je odvisna od parametrov n in m, ki sta poljubni naravni števili (števili prostostnih stopenj).



Slika 3: $\chi^2(1)$ (črna), $\chi^2(2)$ (rjava), $\chi^2(3)$ (rdeča), $\chi^2(4)$ (zelena), $\chi^2(5)$ (rumena)



Slika 4: $S(1)$ (črna), $S(2)$ (rdeča), $S(4)$ (rumena), $N(0,1)$ (zelena)

Literatura

[1] D. Benkovič, Vaje iz biostatistike, Medicinska fakulteta Univerze v Mariboru.

[2] R. Jamnik, Verjetnostni račun in statistika, DMFA, Ljubljana 1995.

[3] J. Sagadin, Statistične metode za pedagoge, Obzorja, Maribor 2003.