# 1 Normalna porazdelitev

Vemo, da je vrednost hemoglobina pri nedopingiranem športniku[1] porazdeljena normalno s povprečjem $\mu = 148$ in varianco $\sigma^2 = 85$.

- Izračunajte verjetnost, da je posameznikova vrednost večja od 166. V ta namen izpeljite formulo:
  - Naj bo $X \sim N(\mu, \sigma^2)$, kako je porazdeljena porazdeljena slučajna spremenljivka $Y = aX + b$, kjer je $a > 0$?

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(aX + b \leq< y) = P(aX \leq y - b) \\ &= P(X \leq \frac{y-b}{a}) = F_X(\frac{y-b}{a}) \\ f_Y(y) &= \frac{1}{a} f_X(\frac{y-b}{a}) \end{aligned}$$

    Za normalno porazdeljeno $X$ torej velja:

$$\begin{aligned} f_Y(y) &= \frac{1}{a \cdot \sqrt{2\pi\sigma^2}} \exp\left\{-\frac{[\frac{y-b}{a} - \mu]^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi(a \cdot \sigma)^2}} \exp\left\{-\frac{[y-(b+a\mu)]^2}{2(a \cdot \sigma)^2}\right\} \end{aligned}$$

    Torej, $Y \sim N(a \cdot \mu + b, (a \cdot \sigma)^2)$.

  - Kaj moramo vzeti kot $a$ in $b$, da bo $Y$ standardizirana normalna spremenljivka.
    $a$ mora biti enak $\frac{1}{\sigma}$, $b$ pa $\frac{-\mu}{\sigma}$. Uporabiti moramo torej transformacijo $Y = \frac{X-\mu}{\sigma}$ in nato verjetnosti odčitati iz tabel za standardizirano normalno porazdelitev (oz. uporabiti ustrezno numerično metodo).

[1] Krvni doping je metoda, pri kateri si športnik kri najprej odvzame, nato pa si jo včrpa pred pomembnim nastopom in tako umetno poveča število rdečih krvničk ter si s tem izboljša trenutno počutje in vzdržljivost. Ker ni udeleženih tujih substanc, krvnega dopinga ni mogoče neposredno odkriti. Zato ga skušajo odkrivati s statističnimi metodami - doping naj bi nakazovale vrednosti krvnih parametrov (hemoglobina), ki pretirano narastejo (včrpanje) oz. padejo (odvzem).

V našem primeru je $X = 166$ in zato $Y = \frac{X-148}{\sqrt{85}} = \frac{166-148}{\sqrt{85}} = 1{,}95$. Iz tabel za standardizirano normalno porazdelitev (ali pa s pomočjo računalnika) izvemo, da je $P(X \leq 166) = P(Y \leq 1{,}95) = 0{,}974$, zato je verjetnost $P(X > 166) = 0{,}026$.

- Izračunajte (simetrične) meje, ki jih nedopingiran športnik preseže z verjetnostjo manj kot 0,01.

  Naj bo $Z$ standardizirana normalna spremenljivka, zanimajo nas meje, izven katerih je vrednost te spremenljivke z verjetnostjo 0,01. Če želimo postaviti simetrične meje, to pomeni, da nas zanimata tisti vrednosti, izven katerih je v repih na vsaki strani verjetnost 0,005. Iz tabel izvemo, da je $P(Z \geq 2{,}57) = 0{,}995$, ustrezna mejna vrednost standardizirane normalne spremenljivke je torej $\pm 2{,}57$.
  $Z = \frac{X-148}{\sqrt{85}}$, zato

$$\begin{aligned} 0{,}995 &= P(\frac{X-148}{\sqrt{85}} \leq 2{,}57) + P(\frac{X-148}{\sqrt{85}} > -2{,}57) \\ &= P(X \leq 148 + 2{,}57 \cdot \sqrt{85}) + P(X > 148 - 2{,}57 \cdot \sqrt{85}) \\ &= P(X \leq 171{,}1) + P(X > 124{,}3) \end{aligned}$$

- Naj bodo meje take, kot ste jih izračunali v prejšnji točki. Športnika testiramo 10x na leto. Kakšna je verjetnost, da vsaj enkrat preseže meje (pri tem predpostavimo, da so meritve narejene v dovolj velikih časovnih presledkih, da so med seboj neodvisne)?

  Naj bo $X$ Bernoullijevo porazdeljena spremenljivka $X \sim Ber(0{,}01)$, kjer je $\{X = 1\} = \{\text{vrednost je izven meja}\}$. Imamo 10 neodvisnih realizacij te slučajne spremenljivke, $X_i$, $i = 1, \ldots, 10$, za vsako velja $P(X_i = 1) = 0{,}01$. Ker so neodvisne, velja $P(X_1 = 0, X_2 = 0, \ldots, X_{10} = 0) = \{P(X_1 = 0)\}^{10}$. Verjetnost, da v 10 meritvah ne preseže meja je torej $P = 0{,}99^{10}$, verjetnost, da jih vsaj enkrat preseže, je $P = 1 - 0{,}99^{10} = 0{,}096$.

- Izračunajte porazdelitev slučajne spremenljivke $X^2$. Katero znano porazdelitev dobite?

$$F_Z(z) \;=\; P(Z \leq z) = P(X^2 \leq z) = P(-\sqrt{z} \leq \sqrt{X} \leq -\sqrt{z})$$

$$
\begin{aligned}
&= F_X(\sqrt{z}) - F_X(-\sqrt{z}) \\
f_Z(z) &= \frac{1}{2\sqrt{z}} f_X(\sqrt{z}) + \frac{1}{2\sqrt{z}} f_X(-\sqrt{z}) \\
&= \frac{1}{2\sqrt{z}}[f_X(\sqrt{z}) + f_X(-\sqrt{z})] \\
&= \frac{1}{2\sqrt{z \cdot 2\pi}}[e^{-z/2} + e^{-z/2}] = \frac{1}{\sqrt{z \cdot 2\pi}} e^{-z/2}
\end{aligned}
$$

Gama porazdelitev ima gostoto $f_T(t) = \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)}$. Če vzamemo, da je $\alpha = \frac{1}{2}$ in $\lambda = \frac{1}{2}$ ter upoštevamo, da je $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, dobimo natanko gornjo formulo. Torej je $X^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2})$ (to je hkrati tudi porazdelitev $\chi^2_1$).

- Raziskovalci na področju športa so dokazali, da je pri biatloncih hemoglobin izven tekmovalnega obdobja porazdeljen kot $N(150, 80)$, med tekmovalnim obdobjem pa kot $N(146, 80)$. Tekmovalno obdobje je pri teh športnikih dolgo približno pol leta. Zanima nas porazdelitev hemoglobina, če ne vemo, kdaj je bil vzorec odvzet. Ali je ta porazdelitev še vedno normalna?

  Definiramo Bernoullijevo porazdeljeno spremenljivko $Y$, ki naj označuje obdobje (0=izven, 1=tekme, verjetnost vsakega izida je 0,5). Poznamo pogojni porazdelitvi:
  $Z|Y = 0 \sim N(150, 80)$, $Z|Y = 1 \sim N(146, 80)$. Porazdelitev Z je torej (uporabimo namig, kjer je $B_1 = \{Y = 0\}$ in $B_2 = \{Y = 1\}$, namesto z verjetnostmi pišemo z gostotami)

$$
\begin{aligned}
f_Z(z) &= f_{Z|Y=0}(z)P(Y = 0) + f_{Z|Y=1}(z)P(Y = 1) \\
&= f_{Z|Y=0}(z)\frac{1}{2} + f_{Z|Y=1}(z)\frac{1}{2} \\
&= \frac{1}{2\sqrt{2\pi 80}} e^{-\frac{(z-146)^2}{2\cdot 80}} [1 + e^{-\frac{16-8(z-146)}{2\cdot 80}}]
\end{aligned}
$$

  Ta spremenljivka v splošnem ni normalno porazdeljena.

# 2 Generiranje slučajnih spremenljivk s pomočjo enakomerne porazdelitve

Generator (psevdo)slučajnih vrednosti iz enakomerne spremenljivke zgenerira željeno število vrednosti $x_i$, ki so porazdeljene kot $X \sim U[0,1]$.

- Kako bi s pomočjo tega generatorja dobili 10 realizacij Bernoullijevo porazdeljene spremenljivke $Y$, pri kateri je $P(Y=1) = 0{,}1$?

  Generiramo[2] 10 vrednosti npr.:

  ```
  > set.seed(4)
  > runif(10)
   [1] 0.585800305 0.008945796 0.293739612 0.277374958
   [5] 0.813574215 0.260427771 0.724405893 0.906092151
   [9] 0.949040221 0.073144469
  ```

  Vrednostim, ki so pod 0,1 damo vrednost 1, ostalim pa 0, torej:

  ```
  > set.seed(4)
  > (runif(10)<0.1)*1
  [1] 0 1 0 0 0 0 0 0 0 1
  ```

- Recimo, da imamo spet 10 enot, vendar jim želimo dati različne verjetnosti, da bodo izžrebane. Prvih pet enot želimo izžrebati z verjetnostjo 0,3, drugih pet pa z verjetnostjo 0,1 (kot primer si zamislimo žreb, v katerem želimo dati prednost ženskam. Verjetnost za vsakega posameznika v našem vzorcu določimo glede na spol - prvih pet je žensk, drugih pet je moških). Kako bi iz istim generatorjem zagotovili ustrezno porazdelitev?

  ```
  > set.seed(4)
  > (runif(10)<c(0.1,0.1,0.1,0.1,0.1,0.3,0.3,0.3,0.3,0.3))*1
   [1] 0 1 0 0 0 1 0 0 0 1
  ```

[2] Kot rešitev vseh praktičnih nalog bo v tem gradivu podana koda za statistični paket R (prostodostopen na `http://cran.r-project.org/`), ki je trenutno med statistiki najbolj razširjen.

- Naj bo $Z = F(X)$, kjer je $F$ porazdelitvena funkcija slučajne spremenljivke $X$.
  - Narišite ustrezen graf (na abscisi so vrednosti $X$, na ordinati pa $Z$)
  - Kakšne vrednosti lahko zavzame spremenljivka $Z$?
    Med 0 in 1
  - Naj bo $X \sim N(0,1)$. Pri kateri vrednosti $X$ bo $Z = 0{,}5$? Kakšna je torej verjetnost, da je $Z \leq 0\ 5$?
    Verjetnost je enaka 0,5
  - Naj bo $X \sim N(0,1)$. Pri kateri vrednosti $X$ bo $Z = 0{,}975$? Kakšna je torej verjetnost, da je $Z \leq 0\ 975$? Verjetnost je enaka 0,975. Vrednosti $Z$ so kvantili porazdelitve $X$.
  - Teoretično izpeljite $F_Z(z)$ za poljuben $F$ (predpostavite, da je $F^{-1}$ definiran za vse vrednosti, ki jih lahko zavzame $X$).

    $$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(F_X(X) \leq z) = P(X \leq F_X^{-1}(z)) \\ &= F_X(F_X^{-1}(z)) = z \end{aligned}$$

    Spremenljivka Z je enakomerno porazdeljena.

- Naj bo $U \sim U[0,1]$ in $X = F^{-1}(U)$. Pokažite, da je $F$ porazdelitvena funkcija spremenljivke $X$.

  $$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

  $F$ je torej kumulativna porazdelitvena funkcija spremenljivke $X$.

- Želimo simulirati vrednosti iz eksponentne porazdelitve ($f(x) = \lambda e^{-\lambda x}$, za $x > 0$). Kako bi jih lahko simulirali z uporabo prej omenjenega generatorja?
  Najprej potrebujemo funkcijo $F$:

  $$\begin{aligned} F_Z(z) &= \int_0^z \lambda e^{-\lambda x} dx \\ &= \frac{\lambda}{-\lambda} e^{-\lambda x}\Big|_0^z \\ &= -1[e^{-\lambda z} - 1] = 1 - e^{-\lambda z} \end{aligned}$$

Inverzna porazdelitev $F^{-1}$ je enaka:

$$\begin{aligned} u &= 1 - e^{-\lambda x} \\ 1 - u &= e^{-\lambda x} \\ -log(1-u) &= \lambda x \\ x &= \frac{-log(1-u)}{\lambda} \end{aligned}$$

Če so vrednosti $u$ torej realizacije enakomerno porazdeljene slučajne spremenljivke $U$, so $x$ realizacije eksponentno porazdeljene spremenljivke $X$.
Kako bi hkrati simulirali vrednosti za posameznike z različno vrednostjo $\lambda$?
Podobno kot zgoraj - le da so vrednosti $\lambda$ lahko različne.