1 Probability

1.1 Normal distribution

The hemoglobin values of an undoped athlete[1] are known to be normally distributed with $\mu = 148$ and variance $\sigma^2 = 85$. Let X denote a measured value of hemoglobin, i.e. $X \sim N(148, 85)$.

- Calculate the probability that the hemoglobin value of an individual exceeds 166. Derive the general formula:
  - Let $X \sim N(\mu, \sigma^2)$, find the distribution of the random variable $Y = aX + b$, where $a > 0$?
    Hint: Start by expressing the distribution function, then derive the density. Can this density be written as a density of a normal covariate?

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

Since X is normal,

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

We get $Y \sim N(a \cdot \mu + b, (a \cdot \sigma)^2)$. Linear transformation of a normal variable remains normal.

[1] In the hope for better performance, the athletes withdraw blood and then reinfuse it before an important competition. In this way, the number of red blood cells is artificially increased, improving the endurance of the athlete. The method is known as blood doping and is highly dangerous and prohibited. Since no foreign substances can be found in the body, there is no way to directly detect the method. Statistics is used to determine unusual values of hemoglobin.

– What do we have to take as a in b, if we want Y to be a standardized normal variable?
a should be equal to $\frac{1}{\sigma}$, b should be equal to $-\frac{\mu}{\sigma}$. The formula $Y = \frac{X-\mu}{\sigma}$ transforms X to the standardized normal scale. We can then use the tables (or the appropriate numerical method) to get the probabilities.
In our case, $X = 166$ and therefore $Y = \frac{X-148}{\sqrt{85}} = \frac{166-148}{\sqrt{85}} = 1.95$. Tables for the standardized normal variable (computer can be used instead) give $P(X \leq 166) = P(Y \leq 1.95) = 0.974$, the probability therefore equals $P(X > 166) = 0.026$.

- Calculate (symmetrical) limits, that are exceeded with probability less than 0.01 by an undoped athlete.

  Let Y be a standardized normal variable. We wish to calculate the limits that encompass 99% of the values. Since we want to have symmetrical limits, the tails are left with probability 0.005 each. From the tables, we learn that $P(Y \geq 2.58) = 0.995$, i.e. the limit value of the standardized normal variable equals ± 2.58.
  $Y = \frac{X-148}{\sqrt{85}}$, therefore

$$\begin{aligned} 0.01 &= P(\frac{X-148}{\sqrt{85}} \leq -2.58) + P(\frac{X-148}{\sqrt{85}} > 2.58) \\ &= P(X \leq 148 - 2.58 \cdot \sqrt{85}) + P(X > 148 + 2.58 \cdot \sqrt{85}) \\ &= P(X \leq 124.2) + P(X > 171.8) \end{aligned}$$

- Consider the calculated limits and let each athlete be tested 10 times in a year. What is the probability that he exceeds the limits at least once (we assume the the athlete is tested within intervals of time that are long enough to ensure no correlation between test results)?

  Let U be a Bernoulli variable $U \sim Ber(0.01)$, where $\{U = 1\} = \{\text{the value is out of limits}\}$. We have 10 independent realizations of this random variable, $U_i,\ i = 1, \ldots, 10$, with $P(U_i = 1) = 0.01$ for each one of them. Since they are independent, $P(U_1 = 0, U_2 = 0, \ldots, U_{10} = 0) = \{P(U_1 = 0)\}^{10}$. The probability that the limits are not exceeded in 10 measurements equals 0.99^{10}, the probability that they are ex-

ceeded at least once, equals $P = 1 - 0.99^{10} = 0.096$.

- Let $Y \sim N(0,1)$. What is the distribution of Y^2?
  Hint: The density of the gamma distribution is given by $f_T(t) = \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)}$.

$$\begin{aligned} F_Z(z) &= P(Z \le z) = P(Y^2 \le z) = P(-\sqrt{z} \le Y \le \sqrt{z}) \\ &= F_Y(\sqrt{z}) - F_Y(-\sqrt{z}) \\ f_Z(z) &= \frac{1}{2\sqrt{z}} f_Y(\sqrt{z}) + \frac{1}{2\sqrt{z}} f_Y(-\sqrt{z}) \\ &= \frac{1}{2\sqrt{z}} [f_Y(\sqrt{z}) + f_Y(-\sqrt{z})] \\ &= \frac{1}{2\sqrt{z \cdot 2\pi}} [e^{-z/2} + e^{-z/2}] = \frac{1}{\sqrt{z \cdot 2\pi}} e^{-z/2} \end{aligned}$$

  The density of the gamma distribution is given by $f_T(t) = \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)}$. If $\alpha = \frac{1}{2}$ and $\lambda = \frac{1}{2}$ ($\Gamma(\frac{1}{2}) = \sqrt{\pi}$), we get exactly the formula above. Therefore, $Y^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2})$ (this is also equal to the χ^2_1 distribution).

- Studies have shown, that the hemoglobin of a biathlon athlete is distributed as $N(150, 80)$ in an out-of-competition phase and as $N(146, 80)$ in the competition phase. Both phases in this sport last for approximately half a year. We wish to know the distribution of hemoglobin if we do not know when the athlete was tested. Is this distribution still normal?
  Hint: $P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$, if $P(\bigcap_{i=1}^n B_i) = 0$ and $P(\bigcup_{i=1}^n B_i) = 1$.

  Define a Bernoulli variable Y, that denotes the phase (0=out-of.competition, 1=competition, probability of each outcome equals 0.5). We know the conditional distributions:
  $Z|Y = 0 \sim N(150, 80)$, $Z|Y = 1 \sim N(146, 80)$. The distribution of Z equals (we use the hint with $B_1 = \{Y = 0\}$ in $B_2 = \{Y = 1\}$, and

plug-in probabilities instead of densities)

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

This variable is not normally distributed.

Understanding the ideas with R:

- Generate 10000 realization of a normal variable $X \sim N(148, 85)$ (`rnorm`). Plot a histogram (`hist`), calculate the proportion of values above 166 (`sum(x>166)/10000`).
- Use the function `qnorm` to find the limits, that an athlete exceeds with probability 0.01. Compare with the proportion in your example.
- Transform the values of X to get a standardized normal variable (`y=(x-148)/sqrt(85)`). Check graphically with a histogram.
- Generate 10 values for 10000 individuals. Calculate the proportion of individuals, that have at least one value outside the interval $[124.2, 171.8]$.
- Plot a histogram of values X^2, compare with the histogram of values generated with functions `rgamma` or `rchisq`.
- Plot the distribution of a random variable described in the last part of the exercise (let the means in the two phases differ more to make sure that the overall distribution is not normal).

1.2 Generating random variables using the uniform distribution

A generator of (pseudo-)random variables results in a number of values x_i, distributed as $X \sim U[0, 1]$.

- How could you use this generator to get 10 Bernoulli distributed variables Y, with $P(Y = 1) = 0.1$?
  We generate 10 values. For example:

```
> 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
```

  The values below 0.1 get the value 1, the rest get the value 0. Therefore:

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

- Say that we again have 10 units, but we wish to assign them different probabilities to be drawn: we want to draw the first five units with the probability 0.3, and the second five with the probablity 0.1 (as an example, say we want to make a random draw, but assign a higher probability to the women, that constitute half of our sample). How can we use the same generator for this purpose?

```
> 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
```

- Let $Z = F(X)$, where F is the distribution function of the random variable X.
  - Sketch the graph of the cumulative distribution function (variable X on the x-axis, and variable Z on the y-axis.)
  - What values can be taken by the variable Z?

    Between 0 and 1

– Let $X \sim N(0,1)$. Which value of X results in $Z = 0.5$? What is the probability that $Z \leq 0.5$?

The probability equals to 0.5

– Let $X \sim N(0,1)$. Which value of X results in $Z = 0.975$? What is the probability that $Z \leq 0.975$?

The probability equals 0.975. The values of Z are the quantiles of the X distribution.

– Express the function $F_Z(z)$ for a given F (assume that the inverse F^{-1} is defined for all possible values of X).

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

The variable Z is uniformly distributed.

- Let $U \sim U[0,1]$ in $X = F^{-1}(U)$. Show that F is the cumulative distribution function of the variable X.
If U is uniform, then $F_U(u) = u$:

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

F is therefore the cumulative distribution function of the variable X.

- We wish to simulate values from the exponential distribution ($f(x) = \lambda e^{-\lambda x}$, za $x > 0$). How can we simulate them using a generator of uniform distribution?
We first need to theoretically express the function F:

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

and its inverse:

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

If the values u are uniformly distributed, the values x shall represent the realizations of an exponentially distributed variable X.

- How would we simulate values for individuals with with different values of λ?
  Same as above, but the values of λ can differ.

Understanding the ideas with R:

- Generate data for 10000 drivers, let 500 of them be drunk and 9500 sober. Let the probability of being involved in a car accident equal 0.3 and 0.003, for a drunk and a sober driver respectively. Calculate the proportion of the accidents on the simulated data and compare it to the actual probability of the accident.

- Generate data for 100 individuals, so that their age is uniformly distributed between 50 and 80. Check graphically with a histogram.

- Say that these individuals have been diagnosed with a lethal disease. Generate the survival times using an exponential distribution, so that the older patients have a higher probability of dying earlier.
  Hint: for example, $\lambda = \text{age}/100$.