Exercise 1:

In this exercise we consider a study of a cohort of nickel smelting workers in South Wales.[1] The data from this study are contained in the data frame `nickel` in the **R** package `Epi`. Once installed, load the package `Epi` by means of the command `library(Epi)`. Then load the data set `nickel` with `data(nickel)`. Make use of `help(nickel)` to become acquainted with the data set.

(a) Use the `Lexis()` function from the package `Epi` to transform the data frame `nickel` into a `Lexis` object. Create a Lexis diagram (for the first 10 rows in the data frame `nickel` only) in which each line in the plot represents the follow-up of a single individual from entry to exit on two time scales: age and calendar time.

(b) Use `points()` to annotate the Lexis diagram with the times of all deaths from lung cancer (International Classification of Diseases (ICD) code 162 or 163).

Exercise 2:

A random variable T with hazard rate

$$h(t) = \lambda\alpha(\lambda t)^{\alpha-1} \quad \text{for} \quad t \geq 0,\ \alpha, \lambda > 0$$

follows a Weibull distribution with parameters α and λ, denoted by $T \sim \mathcal{WB}(\alpha, \lambda)$.

(a) Derive the cumulative hazard function $H(t)$, the survivor function $S(t)$ and probability density function (pdf) $f(t)$.

(b) Set $\alpha = \lambda = 1$ and compute the mean $\mathrm{E}(T)$, variance $\mathrm{Var}(T)$ and the 50th percentile (median lifetime) of the distribution of T. Hint: The 100pth percentile with $p \in [0, 1]$ (also referred to as the pth quantile) of the distribution of T is the smallest t_p so that $S(t_p) \leq 1 - p$, i.e.,

$$t_p = \inf\{t : S(t) \leq 1 - p\} \ .$$

If T is a continuous random variable, then the pth quantile is found by solving the equation $S(t_p) = 1 - p$.

(c) Write a separate function in **R** for $h(t)$, $H(t)$, $S(t)$ and $f(t)$. Plot these functions in **R** for all combinations of parameter values $\alpha = 0.5, 1, 2, 3$ and $\lambda = 0.5, 1, 1.5$.

[1] Breslow, N. E. and Day, N. E (1987): *Statistical Methods in Cancer Research, Volume II - The Design and Analysis of Cohort Studies*, International Agency for Research on Cancer, Lyon.

Exercise 3:

If Y is a random variable that is normally distributed as $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $T := \exp(Y)$ is log-normally distributed, denoted by $T \sim \mathcal{LN}(\mu, \sigma^2)$.

(a) For the random variable T, derive the pdf $f(t)$ and compute the mean $\mathrm{E}(T)$ and variance $\mathrm{Var}(T)$.

(b) Write a separate function in **R** for $h(t)$, $H(t)$, $S(t)$ and $f(t)$. Plot these functions in **R** for $\mu = 0$ and $\sigma^2 = 0.04, 1, 4$. Hint: You may use the functions `dlnorm()` and `plnorm()`.

Exercise 4:

Another quantity of interest in the analysis of time-to-event data is the mean residual lifetime of a random variable T, which is defined as

$$\mathrm{mrl}(t) = \mathrm{E}(T - t \mid T > t) \ .$$

(a) Give an interpretation of $\mathrm{mrl}(t)$.

(b) For a continuous random variable T with density $f(t)$ and survivor function $S(t)$, it can be verified that

$$\mathrm{mrl}(t) = \frac{\int_t^\infty (u-t) f(u)\,\mathrm{d}u}{S(t)} = \frac{\int_t^\infty S(u)\,\mathrm{d}u}{S(t)} \ . \tag{1}$$

Use the expression (1) to compute the mean residual lifetime of an exponentially distributed random variable $T \sim \mathcal{E}(\lambda)$ with $\lambda > 0$. Compare the result to the mean lifetime $\mathrm{E}(T)$.

Exercise 5:

Inverse transform sampling is a method for generating sample numbers at random from a continuous probability distribution with invertible cumulative distribution function (cdf) F. The inverse transform sampling method uses the fact that given a continuous uniform variable U in $[0, 1]$, the random variable $T = F^{-1}(U)$ has cdf F.

(a) Make use of the relationships between the cumulative hazard function $H(t)$ and the cdf $F(t)$ to set up a scheme for generating random numbers from a distribution with invertible $H(t)$.

(b) Write a function in **R** to generate $n \in \mathbb{N}$ random numbers from a distribution with hazard rate $h(t) = t$. First determine $H(t)$ and then use the method developed in (a). Generate $n = 100, 1000, 10000$ random numbers and plot each set of numbers in a separate histogram along with the density $f(t)$.