Exercises

1. **Lack of memory.** Verify that the exponential distribution has the "lack of memory" property, that is, if T is exponentially distributed with parameter $\lambda > 0$ then so is $T - t$ given that $T > t$ for some $t > 0$. How does this fit in with the fact that the hazard function is constant for the exponential distribution ?

2. **The log-normal distribution.** A stochastic variable T is said to be log-normally distributed if $Y = \log(T)$ is normally distributed (with parameters say, μ and σ^2).

   (a) Find the density of T.

   (b) Verify that the log-normal hazard function $h(t)$ has $\lim_{t\to 0} h(t) = 0$ and $\lim_{t\to\infty} h(t) = 0$. Why is this behaviour a restriction for the applicability of the log-normal distribution in survival analysis ?

3. **Mean residual lifetime.** Let T be a continuous random variable with survival function $S(\cdot)$. The mean residual lifetime function $m(\cdot)$ is defined as

$$m(t) = E(T - t|T \geq t).$$

   (a) Prove that

$$m(t) = \frac{\int_t^\infty S(x)dx}{S(t)}.$$

   Also obtain $h(t)$ and hence $S(t)$ in terms of $m(t)$, showing that $m(t)$ uniquely defines the distribution of T (hint: consider $m'(t)$).

   Hint: for a nonnegative continuous resp. discrete random variable X it holds that

$$\mathbb{E}[X] = \int_0^\infty P(X > x)\mathrm{d}x \text{ resp. } \mathbb{E}[X] = \sum_{x=0}^{\infty} P(X \geq x).$$

   (b) Show that specifying a survival time distribution in terms of the hazard, survival, density, or mean residual life time function is equivalent, in the sense that if one function is known, then so are all the other (see also KM page 35).

(c) Show that $m(t) \to \infty$ as $t \to \infty$ for the log-normal distribution.

Hint:

$$\lim_{t\to\infty} m(t) = \lim_{t\to\infty} \left(-\frac{d}{dt} \log f(t) \right)^{-1}$$

where $f(t)$ is the p.d.f. of T. Hint follows by using L'Hospital twice on the left hand side.

4. Show that Type II censoring is a case of independent censoring. That is

$$P(T_1 \in [t, t + \mathrm{d}t[|T_1 \geq t, T_{(r)} \geq t) = P(T_1 \in [t, t + \mathrm{d}t[|T_1 \geq t)$$

Hint: the event $\{T_1 \in [t, t + h[, T_{(r)} > t\}$ is equivalent to the event $\{T_1 \in [t, t + h[, T_l \geq t$ for at least $n - r$ indices l in $\{2, \ldots, n\}\}$.

5. **Likelihood for Type II censoring** Suppose that we observe n individuals with *iid* survival times $T_1, \ldots, T_n$. Recall that in the case of Type II simple censoring, the data consists of the first d order statistics $T_{(1)}, \ldots, T_{(d)}$ where $0 < d \leq n$ is a predetermined number.

Verify that the density for $(T_{(1)}, \ldots, T_{(d)})$ is given by

$$\frac{n!}{(n-d)!} \prod_{i=1}^{d} f(t_{(i)})(1 - F(t_{(d)}))^{n-d}$$

for $t_{(1)} < t_{(2)} < \cdots < t_{(d)}$, where f and F denotes the common density and distribution function for the survival times.

Hint:
Determine the density of $T_{(1)}, \ldots, T_{(d)}$ by inspection of an integral expression for the joint distribution function $P(T_{(1)} \leq t_1, \ldots, T_{(d)} \leq t_d)$. Use the following identity to obtain this expression:

$$P(T_{(1)} \leq t_1, \ldots, T_{(d)} \leq t_d) =$$
$$\binom{n}{d} d! P(T_1 \leq t_1, \ldots, T_d \leq t_d, T_1 < T_2 < \ldots < T_d, T_l > T_d, l = d{+}1, \ldots, n)$$

6. Check that the actuarial estimate reduces to the usual empirical estimate of the survivor function when there is no censoring (i.e. where $P(T \geq t_{u_k})$ is estimated by $\sum_{i=1}^{n} 1(x_i \geq u_k)/n = r(u_k)/n$).

7. **Score function of survival data likelihood** In case of random independent noninformative censoring, the likelihood is of the form

$$\prod_{i=1}^{n} h_X(t_i;\theta)^{\delta_i} S_X(t_i;\theta).$$

Show that the score function of this likelihood has mean zero (assuming usual regularity conditions). That is the so-called first Bartlett identity holds. Hints:

(a) It is enough to consider the case $n = 1$.

(b) One approach: compute

$$\mathbb{E}\frac{\mathrm{d}}{\mathrm{d}\theta} \log \left[f_X(T;\theta)^{\Delta} S_X(T;\theta)^{(1-\Delta)} \right]$$

using $\Delta = 1[X \leq C]$ and $T = \min\{X, C\}$.

(c) Another approach: the score is the derivative of the density $g(t, \delta; \theta)$ for (T, Δ) as specified in the slides for the first lecture. Then use

$$\sum_{\delta=0}^{1} \int_0^{\infty} g(t;\delta;\theta)\mathrm{d}t = 1$$

to evaluate

$$\mathbb{E}\frac{\mathrm{d}}{\mathrm{d}\theta} \log g(T, \Delta; \theta).$$

8. **The delta method.** Suppose that $\sqrt{n}(T_n - c)$ converges in distribution to $N(0, \sigma^2)$ and that g is a continuously differentiable function. Verify that $\sqrt{n}(g(T_n) - g(c))$ converges in distribution to $N(0, \sigma^2(g'(c))^2)$.

Hints: Use the Taylor-expansion

$$g(t) = g(c) + (t - c)g'(c^*)$$

where $|c^* - c| < |t - c|$. Use furthermore that if X_n converges in distribution to X and Y_n converges in probability to a constant y then for any continuous function f on $\mathbb{R}^2$, $f(X_n, Y_n)$ converges in distribution to $f(X, y)$. Also use that if X_n converges in distribution to a constant x then X_n also converges in probability to the constant x.

Comment: if X has a small variance σ^2 around the mean μ we in practice use the delta-method to conclude that $\mathbb{V}\text{ar}g(X)$ is approximately $\sigma^2(g'(\mu))^2$.

Comment: more generally, the following holds: If $\sqrt{n}(T_n - c)$ is a d-dimensional vector converging in distribution to $N_d(0, \Sigma)$ and $g : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable then $\sqrt{n}(g(T_n) - g(c))$ converges in distribution to

$$N(0, \sum_{i,j=1}^{d} \Sigma_{ij} \frac{\partial g}{\partial t_i} \frac{\partial g}{\partial t_j})$$

where Σ_{ij} is the ijth entry in Σ and $\frac{\partial g}{\partial t_i}$ is the ith partial derivative of g.

9. **Greenwood's formula** Give a heuristic derivation of Greenwood's formula by applying the multivariate delta-method and assuming that the $\hat{p}_k = d_k/r(u_{k-1})$ are uncorrelated and $d_k|r(u_{k-1}) \sim \text{bin}(r(u_{k-1}), p_k)$.

10. Show that the Kaplan-Meier estimate reduces to the usual empirical estimate in case of no censoring. Also show that the Greenwood estimate reduces to $\hat{S}(t)(1 - \hat{S}(t))/n$ in case of no censoring agreeing with that $n\hat{S}(t) \sim \text{bin}(n, S(t))$ in case of no censoring.

    Hint: let $t_1^* < t_2^* < \ldots$ denote the ordered death times. If $t_k^* \leq t < t_{k+1}^*$ then $\hat{S}(t) = r(t_{k+1}^*)/n$. Also, with no censoring, $d(t_i^*) = r(t_{i+1}^*) - r(t_i^*)$. Then rewrite the Greenwood estimate as a telescoping sum.

11. Show that the reduced sample estimator (see slides) is unbiased.

12. **Klosterforsikring** T. N. Thiele (1838-1910) was a famous Danish mathematician, astronomist, and statistician. In 1872 he presented computations of the distribution of the time T_G between a birth of a woman and her marriage. This was used in relation with establishment of an annuity for women who never got married and thus needed financial support (details can be found in the post script document on the course web page - it is fun reading).

    Compute the actuarial estimate (and Greenwood standard error) for the "survivor" function $S(t) = P(\text{time for marriage } \geq t), t = 1, 2, 3, \ldots$ (in years) using the Vemmetofte Kloster data (available from the web

page). In the Vemmetofte data, x denotes age, I_x denotes the number of women with age in $[x-1, x[$ who were signed up in the Vemmetofte Kloster, G_x is the number of women who were married in the age $[x-1, x[$, D_x is the number of women who died in the age $[x-1, x[$, U_x is the number of women with age in $[x-1, x[$ who left Vemmetofte Kloster for other reasons than death or marriage, and A_x is the number of women "at risk"/under observation at the beginning of $[x, x+1[$. Thiele computed A_x as

$$A_x = A_{x-1} + I_x - G_x - D_x - U_{x+1},$$

that is, he regarded the U_{x+1} censored women as being censored at time x.

What is the estimated probability that a woman never becomes married ? What is the confidence interval for this estimate ?

Using least squares, Thiele fitted a parametric model given by

$$\log_{10}(S(t)/S(t+1)) = 10^{a-(\log_{10}(t-p)-b)^2/c-10}$$

where he estimated $S(t+1)/S(t)$ by the actuarial estimates of the probability of not getting married in the interval $[t, t+1[$ (note that this is a model for discrete survival data where failures only happen at times $0, 1, 2, \ldots$). He obtained the estimates 8.37476, $b = 1.22588$, $c = 0.104344$, and $p = 8.6363$. Compare Thieles fitted model with your actuarial estimate.

13. We first state an asymptotic result for the Kaplan-Meier estimate (see, e.g. Lawless, J. F., 1982) in the case with random censoring where the censoring times and the survival times have survivor functions $G(\cdot)$ and $S(\cdot)$, respectively.

    Let $\tau < \infty$ satisfy $S(\tau) > 0$, suppose that $1 - S(\cdot)$ is absolutely continuous with density f and that G is continuous. Then the random function

    $$\sqrt{n}(\hat{S}(t) - S(t)),\ 0 < t < \tau$$

    converges weakly (i.e. in distribution) to a mean zero Gaussian process

$(X_t)_{0<t<\tau}$ with covariance function

$$\mathrm{Cov}(X_{t_1}, X_{t_2}) = S(t_1)S(t_2)\int_0^{\min(t_1,t_2)} \frac{f(u)}{S(u)^2 G(u)} du.$$

Assume that the conditions for the asymptotic result are valid. Consider the mean lifetime restricted to τ given by

$$\mu_\tau = \int_0^\tau S(t)dt$$

and the associated estimate

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt$$

where $\hat{S}(\cdot)$ is the Kaplan-Meier estimate.

(a) Show that $\mu_\tau = E\min(T, \tau)$ where T has survival function $S(\cdot)$.

Hint: $\int_0^\tau S(t)dt = \int_0^\tau \int_t^\infty f(u)du$.

(b) Use the asymptotic result for $\hat{S}(\cdot)$ to show that the asymptotic variance of $\sqrt{n}\hat{\mu}_\tau$ is given by

$$\int_0^\tau \frac{A(u)^2 f(u)}{S(u)^2 G(u)} du$$

where $A(u) = \int_u^\tau S(t)dt$.

Hint: Note that asymptotically,

$$\mathrm{Var}(\sqrt{n}\hat{\mu}_\tau) = \int_0^\tau \int_0^\tau \mathrm{Cov}(\sqrt{n}\hat{S}(u), \sqrt{n}\hat{S}(v))dudv.$$

(c) In the special case where there is no censoring let $\tau \to \infty$ and show that $\hat{\mu} = \lim_{\tau\to\infty} \hat{\mu}_\tau$ reduces to the empirical mean of the observed lifetimes and that the asymptotic variance from b. reduces to $\mathrm{Var}(T)$.

14. **Simulation**

    (a) Consider survival data $(T_i, \delta_i) = (\min(X_i, C_i), 1(X_i \leq C_i))$, $i = 1, \ldots, n$, where $X_i \sim E(\lambda)$ and $C_i \sim E(\psi)$ are independent (and $E(\psi)$ denotes the exponential distribution with mean $1/\psi$). For given $\lambda > 0$ and $p \in]0, 1[$ determine ψ so that the probability of censoring equals p.

    (b) Let $n = 100$, $\lambda = 1/5$ and simulate data sets with expected number of censored observations equal to $1, 10, 50, 90, 99$. Compute and plot Kaplan-Meier estimates for the simulated data sets.

    Hint: you may find the R-procedures `rexp`, `survfit`, and `plot.survfit` helpful.

15. Use the R procedures `survfit` and `plot.survfit` to compute and plot survivor functions for the two treatment groups (prednison or placebo) in the cirrhosis data set. Use `survdiff` to test for equal survival in the two groups.

16. **Hypergeometric distribution** Let $X_1 \sim \text{bin}(n_1, p_1)$ and $X_2 \sim \text{bin}(n_2, p_2)$ be independent binomial random variables. Assume $p_1 = p_2 = p$.

    (a) Let $x = x_1 + x_2$ and $n = n_1 + n_2$. Show that

    $$P(X_1 = x_1 | X_1 + X_2 = x) = \frac{\binom{n_1}{x_1}\binom{n-n_1}{x-x_1}}{\binom{n}{x}},$$

    i.e. the conditional distribution of X_1 given the sum is a hypergeometric distribution (that does *not* depend on p).

    (b) Show that the mean and the variance of the above hypergeometric distribution are

    $$n_1 \frac{x}{n} \text{ and } \frac{n_1 x(n - n_1)(n - x)}{n^2(n - 1)}$$

17. Suppose that T has hazard function

    $$\lambda(t) = \lambda_0(t) \exp(z\beta)$$

    and that g is a strictly increasing and continuously differentiable function. Show that $\tilde{T} = g(T)$ has hazard function

    $$\tilde{\lambda}(t) = \lambda_0(g^{-1}(t)) \exp(z\beta)/g'(g^{-1}(t)).$$

18. Suppose that $T_1, \ldots, T_n$ are independent continuous survival times with hazard rates λ_i, $i = 1, \ldots, n$. Let $T = \min(T_1, \ldots, T_n)$ and verify that the hazard rate of T is given by $\lambda(t) = \sum_{i=1}^{n} \lambda_i(t)$.

    Hint: use that the survivor function for T is given by

$$P(T \geq t) = \prod_{i=1}^{n} P(T_i \geq t).$$

19. **log rank vs. score test.** Consider the Cox's proportional hazards model in the case where the covariates are of the form

$$z_i = \begin{cases} 1 & \text{if } i\text{th individual got treatment no. 1} \\ 0 & \text{if } i\text{th individual got treatment no. 2} \end{cases}$$

    The score test statistic for no treatment effect is given by

$$u(0)^2 / j(0)$$

    where $u(\beta)$ and $-j(\beta)$ are the first and second derivatives of the log Cox's partial likelihood. Assume that there are no tied death times and check that the score test statistic coincides with the log rank statistic.

20. **Poisson process.** Consider a counting process $N(t)$, $t \geq 0$, and let $T_1, T_2, \ldots$ denote the jump times of $N(t)$, $t \geq 0$. Let $W_i = T_i - T_{i-1}$, $i = 1, 2, \ldots$ be the waiting times between the jump times (letting $T_0 = 0$). Then $N(t)$, $t \geq 0$, is Poisson process with rate $\lambda > 0$ if and only if

    C1 The W_i are independent and exponentially distributed with rate λ (i.e. mean $1/\lambda$).

    Suppose $N(t)$ is a Poisson process. Show that for any $t > 0$, $N(t)$ is Poisson distributed with mean λt. Actually one can show that C1 is equivalent to the following to conditions:

    C2 For $0 < s < t$, $N(t) - N(s)$ is Poisson distributed with mean $\lambda(t - s)$.

    C3 For $0 < s_1 < t_1 < s_2 < t_2$, $N(t_1) - N(s_1)$ and $N(t_2) - N(s_2)$ are independent (i.e. the Poisson process has independent increments).

Suppose that $N(t)$, $t \geq 0$, is a Poisson process with rate λ. Show that $M(t) = N(t) - \lambda t$, $t \geq 0$, is a martingale (i.e. show that $E[M(t)|N(s)] = M(s)$ for $0 \leq s < t$).

21. Consider the actuarial estimate in the case where there is no censoring and recall that $\hat{p}_i = 1 - \hat{q}_i = (n_i - d_i)/n_i$ where n_i is the number of persons alive just before time t_i and d_i is the number of deaths in $[t_i, t_{i+1}[$. Show by counter example that $\hat{p}_i$ and $\hat{p}_j$ are not independent even though $\mathrm{Cov}(\hat{p}_i, \hat{p}_j) = 0$ (hint: consider e.g. $\hat{p}_0$ and $\hat{p}_1$).

22. **Interval censoring** Let T be a survival time with the exponential distribution $\mathrm{Exp}(\lambda)$ where $\lambda = ET = 3$. Assume that T is interval censored so that we only observe

$$X = \begin{cases} 1 & \text{if } T \leq 1, \\ T & \text{if } 1 < T < 3, \\ 3 & \text{if } 3 \leq T. \end{cases}$$

    (a) Plot the distribution function of X.

    (b) Find the density of X (with respect to the sum of Lebesgue and counting measure on $\{1, 3\}$), i.e. a function $f : \mathbb{R} \to \mathbb{R}_+$ so that

$$P(X \leq x) = \int_0^x f(z)dz + f(1)1(1 \leq x) + f(3)1(3 \leq x).$$

23. Suppose that the survival times $T_1, \ldots, T_n$ follow Cox's proportional hazards model and there is no censoring. Verify that the likelihood based on the ranks $R_1, \ldots, R_n$,

$$P(R_1 = r_1, \ldots, R_n = r_n) = P(T_{r_1} < \ldots < T_{r_n})$$

is Cox's partial likelihood.

Hint: write down the integral for the right hand side and solve it 'backwards' by the method of substitution.

More elaborate hint: assume for ease of notation and without loss of generality that $r_i = i$, $i = 1, \ldots, n$. Then the partial likelihood is

$$\prod_{i=1}^{n} \frac{\exp(z_i^{\mathsf{T}}\beta)}{\sum_{j=i}^{n} \exp(z_j^{\mathsf{T}}\beta)} = \prod_{i=1}^{n} \frac{1}{a_i}$$

where $a_n = 1$ and $a_i = \sum_{j=i+1}^{n} \exp((z_j - z_i)^\mathsf{T}\beta) + 1$, $i = 1, \ldots, n-1$. Also note that under the proportional hazards model,

$$S_{l+1}(u) = S_l(u)^{\exp((z_{l+1}-z_l)^\mathsf{T}\beta)}.$$

One can now show that (using integration by substitution)

$$\int_v^\infty f_l(u) S_{l+1}(u)^{a_{l+1}} \mathrm{d}u = \frac{1}{a_l} S_l(v)^{a_l}.$$

This can be used for the stepwise solution of the integral.

24. A Brownian motion $\{B(s)\}_{s\geq 0}$ is a continuous-time zero-mean Gaussian process[1] with $B(0) = 0$ and $\mathbb{C}\mathrm{ov}(B(s), B(t)) = \min(t, s)$ for $s, t \geq 0$.

    - Show that a Brownian motion has uncorrelated and hence independent increments over disjoint intervals.
    - show that a Brownian motion is a martingale with respect to its own history:

    $$\mathbb{E}[B(t)|B(u), 0 \leq u \leq s] = B(s).$$

25. Show heuristically that if M is a martingale and K is a predictable process (both with respect to $(\mathcal{F}_t)_{t\geq 0}$) then

    (a) $\tilde{M}(t) = \int_0^t K(u)\mathrm{d}M(u)$ is a martingale.

    (b) $\tilde{M}$ has predictable variation process $< \tilde{M} > (t) = \int_0^t K(u)^2 \mathrm{d} < M > (u)$.

26. Show that a martingale has uncorrelated increments (over disjoint intervals).

27. Show that using the timedependent covariate $z_i(t) = a_i + bt$ for the ith subject in a Cox regression is the same as using age a_i at $t = 0$ as a fixed covariate.

[1] I.e. all finite-dimensional distributions are Gaussian