NTNU
Norwegian University of Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4275 Lifetime Analysis**

Academic contact during examination: Håkon Tjelmeland

Phone:

Examination date: June 2nd 2023 – solution sketch

Examination time (from–to): 09.00-13.00

Permitted examination support material: C:

Tabeller og formler i statistikk, Akademika
A yellow sheet of paper (A5 with a stamp) with personal handwritten formulas and notes
A specific basic calculator

Other information:

Language: English

Number of pages: 9

Number of pages enclosed: 0

Checked by:

Informasjon om trykking av eksamensoppgave	
Originalen er:	
1-sidig ☐	**2-sidig** ☒
sort/hvit ☒	**farger** ☐
skal ha flervalgskjema ☐	

Date Signature

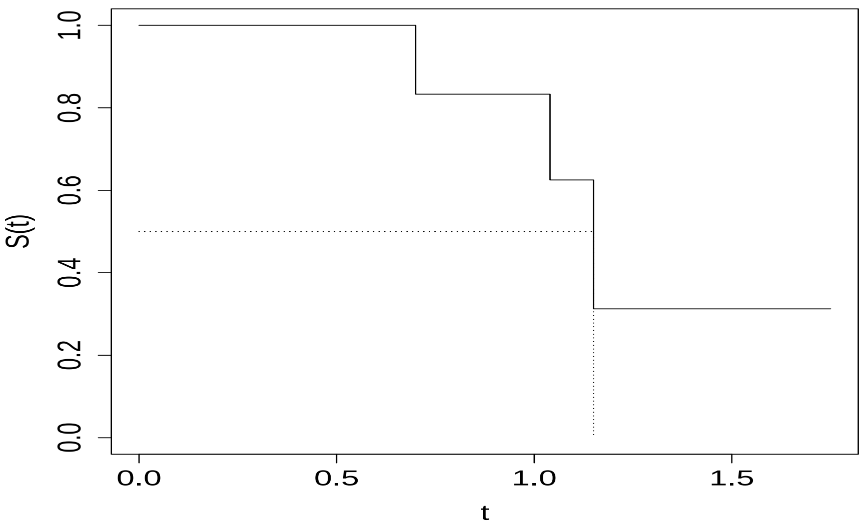

Figure 1: Estimated survival function $\widehat{S}(t)$ for $t \in [0, 1.75]$. How to estimate the median survival time is illustrated with a dotted line.

Problem 1

The Kaplan-Meier estimator is

$$\widehat{S}(t) = \prod_{j:T_j \leq t} \left(1 - \frac{1}{Y(T_j)}\right). \tag{1}$$

In the data set we have only three observed survival times, namely 0.70, 1.04 and 1.15. The number of individuals under risk just prior to each of these three times are $Y(0.70) = 6$, $Y(1.04) = 4$ and $Y(1.15) = 2$, respectively. Thereby, for $t < 0.70$ we have $\widehat{S}(t) = 1$, for $t \in [0.70, 1.04)$ we have $\widehat{S}(t) = 1 - \frac{1}{6} = \frac{5}{6} = 0.833$, for $t \in [1.04, 1.15)$ we have $\widehat{S}(t) = \frac{5}{6} \cdot \left(1 - \frac{1}{4}\right) = 0.625$, and for $t \geq 1.15$ we have $\widehat{S}(t) = 0.625 \cdot \left(1 - \frac{1}{2}\right) = 0.3125$.

A plot of $\widehat{S}(t)$ is given in Figure 1.

The median survival time can estimated from $\widehat{S}(t)$ as illustrated in Figure 1, and here the estimated median survival time becomes 1.15.

Problem 2

We start by finding the cumulativ distribution function,

$$\begin{aligned}
F(t) &= \int_0^t f(u)du = \int_0^t \varphi e^{-\theta u} \exp\left\{-\frac{\varphi}{\theta}\left(1-e^{\theta u}\right)\right\} du \\
&= \left[-\exp\left\{-\frac{\varphi}{\theta}\left(1-e^{-\theta u}\right)\right\}\right]_{u=0}^{u=t} \\
&= -\exp\left\{-\frac{\varphi}{\theta}\left(1-e^{-\theta t}\right)\right\} - (-1) \\
&= 1-\exp\left\{-\frac{\varphi}{\theta}\left(1-e^{-\theta t}\right)\right\}.
\end{aligned}$$

The survival function thereby becomes

$$\underline{\underline{S(t)}} = 1 - F(t) = \underline{\underline{\exp\left\{-\frac{\varphi}{\theta}\left(1-e^{-\theta t}\right)\right\}}}.$$

Since we have that $S(t) = e^{-A(t)}$, this in turn gives that the integrated hazard rate is

$$A(t) = -\ln(S(t)) = -\left(-\frac{\varphi}{\theta}\left(1-e^{-\theta t}\right)\right) = \frac{\varphi}{\theta}\left(1-e^{-\theta t}\right),$$

which gives the hazard rate

$$\underline{\underline{\alpha(t)}} = A'(t) = \frac{\varphi}{\theta}\cdot(-e^{-\theta t})\cdot(-\theta) = \underline{\underline{\varphi e^{-\theta t}}}.$$

Problem 3

From the problem text we have that

$$N(t) = \int_0^t \lambda(s)ds + M(t) = \int_0^t Y(s)\alpha(s)ds + M(t),$$

which in incremental form becomes

$$dN(t) = Y(t)\alpha(t)dt + dM(t).$$

Inserting this into the expression for the Nelson-Aalen estimator we get

$$\begin{aligned}
\widehat{A}(t) &= \int_0^t \frac{J(s)}{Y(s)}\left(Y(s)\alpha(s)ds + dM(s)\right) \\
&= \int_0^t J(s)\alpha(s)ds + \int_0^t \frac{J(s)}{Y(s)}dM(s) \\
&= A^\star(t) + \int_0^t \frac{J(s)}{Y(s)}dM(s),
\end{aligned}$$

where we in the last transition have used how $A^\star(t)$ is defined in the problem text. Thus, we have

$$\underline{\underline{\widehat{A}(t) - A^\star(t) = \int_0^t \frac{J(s)}{Y(s)} dM(s),}} \tag{2}$$

as we should show. The $Y(s)$ is by definition a predictable process, and since $J(s) = I(Y(s) > 0)$ is just a function of $Y(s)$ also $J(s)$ becomes a predictable process. Thereby also the integrand above, $J(s)/Y(s)$, is a predictable process, so the right hand side of (2) is a stochastic integral with respect to a mean zero martingale. As we know that a stochastic integral with respect to a mean zero martingale is itself a mean zero martingale, we thereby have that $\widehat{A}(t) - A^\star(t)$ is a mean zero martingale.

Using (2) and computational rules for the optional variation process of a stochastic integral we get

$$\begin{aligned}
\left[\widehat{A} - A^\star\right](t) &= \left[\int \frac{J}{Y} dM\right](t) \\
&= \int_0^t \left(\frac{J(s)}{Y(s)}\right)^2 d[M](s) \\
&= \int_0^t \frac{J(s)^2}{Y(s)^2} d[M](s).
\end{aligned}$$

Since $J(s)$ is an indicator function we have that $J(s)^2 = J(s)$. Moreover, for counting process martingales we know that $[M](s) = N(s)$ so we get that

$$\underline{\underline{\left[\widehat{A} - A^\star\right](t) = \int_0^t \frac{J(s)}{Y(s)^2} dN(s).}} \tag{3}$$

Since the right hand side of (3) is an integral with respect to a counting process, the integral just becomes a sum over the times where the counting process jumps. Thus, by using that whenever $dN(s) = 1$ we must have that $J(s) = 1$, we get

$$\left[\widehat{A} - A^\star\right](t) = \sum_{j:T_j \le t} \frac{1}{Y(s)^2} = \widehat{\sigma}^2(s). \tag{4}$$

We generally know that the variance of a mean zero martingale at any time t equals the expected value of the associated optional variation process at the same time t. Combining this result for $\widehat{A}(t) - A^\star(t)$ with (4) we get

$$\mathrm{Var}\left(\widehat{A}(t) - A^\star(t)\right) = \mathrm{E}\left(\left[\widehat{A} - A^\star\right](t)\right) = \mathrm{E}\left(\widehat{\sigma}^2(t)\right).$$

This equation says that $\widehat{\sigma}^2(t)$ is an unbiased estimator for the variance of $\widehat{A}(t) - A^\star(t)$, which is exactly what we should show.

Problem 4

a) Taking the logarithm of the general expression for the partial likelihood given in the problem text, and using that the covariates are time invariant and that β and x_ℓ are vectors of size two in our situation, we get

$$\begin{aligned}\ell(\beta_1,\beta_2) &= \sum_j \left[\beta^T x_{i_j} - \ln\left(\sum_{\ell\in\mathcal{R}_j} e^{\beta^T x_\ell}\right)\right]\\ &= \sum_j \left[\beta_1 x_{i_j1} + \beta_2 x_{i_j2}\right] - \sum_j \ln\left(\sum_{\ell\in\mathcal{R}_j} \exp\left\{\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2}\right\}\right).\end{aligned}$$

Since components that fails is immediately repaired we have that all components are under risk for failure at all times, so $\mathcal{R}_j = \{1,\ldots,n\}$ for all j. This gives that

$$\begin{aligned}\ell(\beta_1,\beta_2) &= \sum_j \left[\beta_1 x_{i_j1} + \beta_2 x_{i_j2}\right] - \sum_j \ln\left(\sum_{\ell=1}^n \exp\left\{\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2}\right\}\right)\\ &= \sum_j \left[\beta_1 x_{i_j1} + \beta_2 x_{i_j2}\right] - m\ln\left(\sum_{\ell=1}^n \exp\left\{\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2}\right\}\right),\end{aligned}$$

where m is the total number of observed failures, and we in the last transition have used that both covariates are time invariant. Next, we use that $x_{\ell 1} \in \{0,1\}$ to split each of the two sums above into a sum of two sums,

$$\begin{aligned}\ell(\beta_1,\beta_2) = &\sum_{j:x_{i_j1}=0} \beta_2 x_{i_j2} + \sum_{j:x_{i_j1}=1}\left[\beta_1 + \beta_2 x_{i_j2}\right]\\ &- m\ln\left(\sum_{\ell:x_{\ell 1}=0}\exp\left\{\beta_2 x_{\ell 2}\right\} + \sum_{\ell:x_{\ell 1}=1}\exp\left\{\beta_1 + \beta_2 x_{\ell 2}\right\}\right)\\ &= m^{(1)}\beta_1 + \beta_2\sum_j x_{i_j2} - m\ln\left(\sum_{\ell:x_{\ell 1}=0}\exp\left\{\beta_2 x_{\ell 2}\right\} + e^{\beta_1}\sum_{\ell:x_{\ell 1}=1}\exp\left\{\beta_2 x_{\ell 2}\right\}\right)\\ &= m^{(1)}\beta_1 + \beta_2\sum_j x_{i_j2} - m\ln\left(h_0(\beta_2) + e^{\beta_1}h_1(\beta_2)\right),\end{aligned}$$

where $m^{(1)} = \sum_{j:x_{i_j1}=1} 1$ is the number of failures for components from manufacturer B, and $h_0(\beta_2)$ and $h_1(\beta_2)$ are as given in the problem text. This is exactly the expression we were asked to show.

b) To try to find analytical expressions for the maximum partial likelihood estimators we start to find expressions for the partial derivatives. We start with the partial derivative with respect to β_1,

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_1} &= m^{(1)} - m \cdot \frac{e^{\beta_1} h_1(\beta_2)}{h_0(\beta_2) + e^{\beta_1} h_1(\beta_2)} \\ &= m^{(1)} - m \cdot \frac{h_1(\beta_2)}{e^{-\beta_1} h_0(\beta_2) + h_1(\beta_2)}.\end{aligned}$$

We see that we easily can solve $\frac{\partial \ell}{\partial \beta_1} = 0$ with respect to $e^{-\beta_1}$, and thereby also with respect to β_1. We get

$$\begin{aligned}e^{-\beta} &= \frac{\frac{h_1(\beta_2)}{m^{(1)}}}{\frac{h_0(\beta_2)}{m^{(0)}}}, \\ \beta_1 &= \ln\left(\frac{h_0(\beta_2)}{m^{(0)}}\right) - \ln\left(\frac{h_1(\beta_2)}{m^{(1)}}\right), \end{aligned} \tag{5}$$

where $m^{(0)} = m - m^{(1)}$ is the number of failures for components from manufacturer A. Thereby we have a corresponding relation between the estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$,

$$\widehat{\beta}_1 = \ln\left(\frac{h_0(\widehat{\beta}_2)}{m^{(0)}}\right) - \ln\left(\frac{h_1(\widehat{\beta}_2)}{m^{(1)}}\right).$$

The partial derivative of the log partial likelihood function with respect to β_2 becomes

$$\frac{\partial \ell}{\partial \beta_2} = \sum_j x_{i_j 2} - m \cdot \frac{h_0'(\beta_2) + e^{\beta_1} h_1'(\beta_2)}{h_0(\beta_2) + e^{\beta_1} h_1(\beta_2)}, \tag{6}$$

where

$$\begin{aligned}h_0'(\beta_2) &= \sum_{\ell : x_{\ell 1} = 0} x_{\ell 2} \exp\{\beta_2 x_{\ell 2}\}, \\ h_1'(\beta_2) &= \sum_{\ell : x_{\ell 1} = 1} x_{\ell 2} \exp\{\beta_2 x_{\ell 2}\}.\end{aligned}$$

When inserting the expressions for $h_0'(\beta_1)$ and $h_1'(\beta_2)$ into (6) we see that there is no way of getting β_2 outside the various sums. Thereby it is not possible to solve $\frac{\partial \ell}{\partial \beta_2} = 0$ with respect to β_2, and this doesn't change if we insert the expression we found above for β_1 as a function of β_2.

To find $\widehat{\beta}_2$ we therefore need to optimise numerically the resulting log profile partial likelihood for β_2, which we get by inserting (5) into the expression for the log-likelihood. The log profile partial likelihood for β_2 thus becomes

$$\begin{aligned}
\ell_p(\beta_2) &= m^{(1)} \ln\left(\frac{h_0(\beta_2)}{m^{(0)}}\right) - m \ln\left(\frac{h_1(\beta_2)}{m^{(1)}}\right) + \beta_2 \sum_j x_{i_j 2} - m \ln\left(h_0(\beta_2) + \frac{\frac{h_0(\beta_2)}{m^{(0)}}}{\frac{h_1(\beta_2)}{m^{(1)}}} \cdot h_1(\beta_2)\right) \\
&= \text{const} + m^{(1)} \ln(h_0(\beta_2)) - m \ln(h_1(\beta_2)) + \beta_2 \sum_j x_{i_j 2} - m \ln(h_0(\beta_2)) \\
&= \text{const} + \beta_2 \sum_j x_{i_j 2} - m^{(0)} \ln(h_0(\beta_2)) - m^{(1)} \ln(h_1(\beta_2)),
\end{aligned}$$

where const is a term that is constant as a function of β_2.

Problem 5

a) When $\text{age}_i = 45$ and $\text{sexmale}_i = 0$ the estimated intensity process is

$$\lambda_i(t) = \widehat{b} t^{\widehat{k}-1} \exp\left\{\widehat{\beta}_1 \cdot 45 + \widehat{\beta}_2 \cdot 0\right\}.$$

From the R output we find the parameter estimates

$$\begin{aligned}
\widehat{\mu} &= 11.24975, \\
\widehat{\gamma}_1 &= -0.03995, \\
\widehat{\gamma}_2 &= -0.16387, \\
\widehat{\tau} &= -0.33389.
\end{aligned}$$

Transforming to the other parameterisation we get the estimates

$$\begin{aligned}
\widehat{b} &= \exp\left\{-\left[\widehat{\tau} + \widehat{\mu} e^{-\widehat{\tau}}\right]\right\} = 2.102133 \cdot 10^{-7}, \\
\widehat{\beta}_1 &= -\widehat{\gamma}_1 e^{-\widehat{\tau}} = 0.05578576, \\
\widehat{\beta}_2 &= -\widehat{\gamma}_2 e^{-\widehat{\tau}} = 0.2288264, \\
\widehat{k} &= e^{-\widehat{\tau}} = 1.39639.
\end{aligned}$$

So the estimated intensity process becomes

$$\begin{aligned}
\underline{\underline{\lambda_i(t)}} &= 2.102133 \cdot 10^{-7} \cdot t^{0.39639} \exp\{0.05578576 \cdot 45 + 0.2288264 \cdot 0\} \\
&\underline{\underline{= 2.587589 \cdot 10^{-6} \cdot t^{0.39639}}}.
\end{aligned}$$

To find a 95% confidence interval for k we first find a 95% confidence interval for τ. As we know that the vector of parameter estimators is approximately multivariate normal distributed, it follows that the estimator $\widehat{\tau}$ has approximately a univariate normal distribution. From the R output we find that the standard deviation of $\widehat{\tau}$ is estimated to $\widehat{\mathrm{SD}}[\widehat{\tau}] = 0.05650$. So thereby a 95% confidence interval for τ becomes

$$\left[\widehat{\tau} - z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}], \widehat{\tau} + z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}\right] = [-0.44463, -0.22315],$$

where we have used that $z_{0.025} = 1.96$. This implies that

$$P\left(\widehat{\tau} - z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}] \le \tau \le \widehat{\tau} + z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}]\right) \approx 0.95,$$

which, using the relation $k = e^{-\tau} \Leftrightarrow \tau = -\ln(k)$, gives that

$$\begin{aligned}
&P\left(\widehat{\tau} - z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}] \le -\ln(k) \le \widehat{\tau} + z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}]\right) \approx 0.95 \\
&P\left(-\left(\widehat{\tau} + z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}]\right) \le \ln(k) \le -\left(\widehat{\tau} - z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}]\right)\right) \approx 0.95 \\
&P\left(\exp\left\{-\left(\widehat{\tau} + z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}]\right)\right\} \le k \le \exp\left\{-\left(\widehat{\tau} - z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}]\right)\right\}\right) \approx 0.95
\end{aligned}$$

So thereby a 95% confidence interval for k is

$$\begin{aligned}
&\left[\exp\left\{-\left(\widehat{\tau} + z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}]\right)\right\}, \exp\left\{-\left(\widehat{\tau} - z_{0.025}\widehat{\mathrm{SD}}[\widehat{\tau}]\right)\right\}\right] \\
&= \left[\exp\left\{-(-0.22315)\right\}, \exp\left\{-(-0.44463)\right\}\right] \\
&= \underline{\underline{[1.250008, 1.559913]}}.
\end{aligned}$$

In an exponential regression model one would have $k = 1$. As the confidence interval for k does not include the value 1, one should not expect the exponential regression model to give a good fit to this dataset. Alternatively one can see the same directly from the parameter esimates given by R. In the R parameterisation, an exponential regression model would have $\tau = 0$, and from the *R* output we see that the τ parameter is significantly different from zero (with a p-value of $3.4 \cdot 10^{-9}$).

b) When $\text{age}_i = 50$ and $\text{sexmale}_i = 1$ the estimated intensity process is

$$\begin{aligned}
\lambda_i(t) &= \widehat{b}t^{\widehat{k}-1}\exp\{\widehat{\beta}_1 \cdot 50 + \widehat{\beta}_2 \cdot 1\} \qquad (7)\\
&= 2.102133 \cdot 10^{-7} \cdot t^{1.39639-1}\exp\{0.05578576 \cdot 50 + 0.2288264\} \\
&= 4.299427 \cdot 10^{-6} \cdot t^{1.39639-1}.
\end{aligned}$$

The corresponding integrated intensity process becomes

$$\begin{aligned}
\Lambda_i(t) &= \int_0^t \lambda_i(u)du = \frac{4.299427 \cdot 10^{-6}}{1.39639} \cdot t^{1.39639} \\
&= 3.078959 \cdot 10^{-6} \cdot t^{1.39639}.
\end{aligned}$$

The corresponding (estimated) survival function becomes

$$\begin{aligned}\widehat{S}_i(t) &= \exp\{-\Lambda_i(t)\}\\ &= \underline{\underline{\exp\{-3.078959\cdot 10^{-6}t^{1.39639}\}}}.\end{aligned}$$

In particular we have

$$\underline{\underline{\widehat{S}_i(10\,000)}} = \exp\{-3.078959\cdot 10^{-6}10000^{1.39639}\} = \underline{\underline{0.3055413}}.$$

To find a 95% confidence interval for $S(10\,000)$ we can first combine (7) with the given expressions for b, β_1, β_2 and k as functions of μ, γ_1, γ_2 and τ to find an expression for $S(10\,000)$ as a function of μ, γ_1, γ_2 and τ,

$$S(10\,000) = g(\mu,\gamma_1,\gamma_2,\tau)$$

say. Correspondingly we then have

$$\widehat{S}(10\,000) = g(\widehat{\mu},\widehat{\gamma}_1,\widehat{\gamma}_2,\widehat{\tau}).$$

We can then do a Taylor series expansion of g around the true (unkown) parameter values μ, γ_1, γ_2 and τ, and make an approximation by including only the zero and first order terms. So we then get

$$\begin{aligned}\widehat{S}(10\,000) = g(\widehat{\mu},\widehat{\gamma}_1,\widehat{\gamma}_2,\widehat{\tau}) &\approx g(\mu,\gamma_1,\gamma_2,\tau) + \frac{\partial g}{\partial \mu}(\mu,\gamma_1,\gamma_2,\tau)(\widehat{\mu}-\mu)\\ &+ \frac{\partial g}{\partial \gamma_1}(\mu,\gamma_1,\gamma_2,\tau)(\widehat{\gamma}_1-\gamma_1) + \frac{\partial g}{\partial \gamma_2}(\mu,\gamma_1,\gamma_2,\tau)(\widehat{\gamma}_2-\gamma_2)\\ &+ \frac{\partial g}{\partial \tau}(\mu,\gamma_1,\gamma_2,\tau)(\widehat{\tau}-\tau),\end{aligned}$$

which is a linear function of the vector $(\widehat{\mu},\widehat{\gamma}_1,\widehat{\gamma}_2,\widehat{\tau})$. We know from general theory that $(\widehat{\mu},\widehat{\gamma}_1,\widehat{\gamma}_2,\widehat{\tau})$ is approximately multivariate normal with mean vector $(\mu,\gamma_1,\gamma_2,\tau)$ and a covariance matrix for which we have an estimate given in the R output. We thereby have that $\widehat{S}(10\,000)$ is approximately normal. Using the linearised expression above we find that

$$\mathrm{E}[\widehat{S}(10\,000)] = g(\mu,\gamma_1,\gamma_2,\tau) = S(10\,000).$$

We can correspondingly find an expression for the variance, which will be a function of the unknown parameters via the expressions for the partial derivatives and a function of the covariance matrix for $(\widehat{\mu},\widehat{\gamma}_1,\widehat{\gamma}_2,\widehat{\tau})$. We can then find an estimate of the variance by, in the expression for the variance of $\widehat{S}(10\,000)$, plugging in the estimates $\widehat{\mu},\widehat{\gamma}_1,\widehat{\gamma}_2,\widehat{\tau}$ for $\mu,\gamma_1,\gamma_2,\gamma_2$ and using the estimated covariance matrix for $(\widehat{\mu}, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\tau})$ provided by R. Thus, we now have that

$$\widehat{S}(10\,000) \approx N(S(10\,000), \sigma^2),$$

where σ^2 is the value we found as an estimate for the variance of $\widehat{S}(10\,000)$. Finally from this we easily find the confidence interval for $S(10\,000)$ as

$$\left[\widehat{S}(10\,000) - z_{0.025}\sigma, \widehat{S}(10\,000) + z_{0.025}\sigma\right].$$