Applied statistics: Coursework 1

Henry Haustein

26th February 2019

Contents

1 Task 1

1.1 Part (1)

In the given data were two out of 26 data points with an Al/Be ratio of more than 4.5. That means

$$\hat{p} = \frac{2}{26} = \frac{1}{13}$$

1.2 Part (2)

Using the following formula from the lecture we get the 95% confidence interval:

$$\begin{gathered}\hat{p} \pm 2 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ \frac{1}{13} \pm \underbrace{2 \cdot \sqrt{\frac{\frac{1}{13} \cdot \frac{12}{13}}{26}}}_{0.1045}\end{gathered}$$

Our 95% confidence interval is [-0.0276,0.1814] which means that we are 95% sure that the true proportion lies between -0.0276 and 0.1814.

1.3 Part (3)

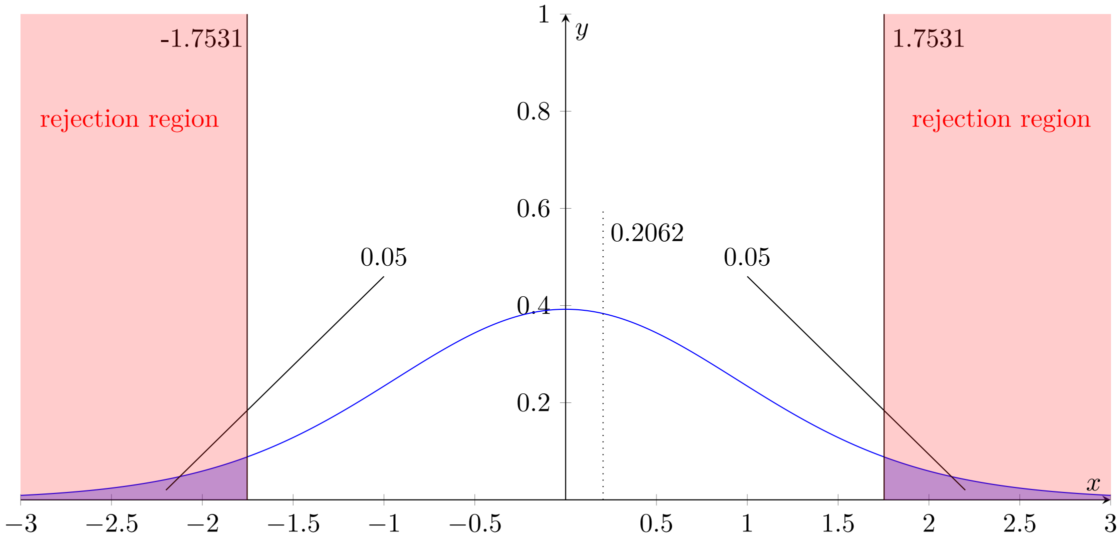
1.4 Part (4)

2 Task 2

2.1 Part (1)

```
x = [-4.5, -1, -0.5, -0.15, 0, 0.01, 0.02, 0.05, ...
0.15, 0.2, 0.5, 0.5, 1, 2, 3];
m = mean(x);
s = std(x);
```

null hypothesis	H_0: $\mu = 0$
alternative hypothesis	H_A: $\mu \neq 0$
t-test for μ	$t = \frac{m-0}{\frac{s}{\sqrt{15}}} = \frac{0.0853}{\frac{1.6031}{\sqrt{15}}} = 0.2062$
rejection region	`tinv(0.05,15)` = -1.7531
conclusion	t lies not in the rejection region so H_0 is accepted at the 10% significance level.



2.2 Part (2)

If we reduce the significance level our rejection region gets smaller. With $\alpha = 0.05$ the rejection region will start at `tinv(0.025,15)` = -2.1314. The t calculated in part (1) won't change $\Rightarrow$ our decision won't change too.

To get the type 2 error we use the MATLAB function `sampsizepwr` and $\mathit{type\,2\,error} = 1 - \mathit{power}$.

```
testtype = 't';
p0 = [0 1.6031];
p1 = 0.0853;
n = 15;
power = sampsizepwr(testtype,p0,p1,[],n)
```

This gives $power = 0.0542 \Rightarrow type\,2\,error = 0.9458$. This is the probability of wrongly accepting H_0 when it is false.

2.3 Part (3)

3 Task 3

3.1 Part (1)

First of all we need to prepare the data:

```
raw = load('input_data.txt');
data = reshape(raw,[1 500]); %produce a single vector
```

After that we do for every distribution (normal, exponential, uniform, lognormal, Rayleigh, gamma) the same procedure:

1. Estimate the parameter. This is often done with the function `<distribution>fit` but for estimating the parameters in the gamma distribution I used `fitdist(data', 'Gamma')` because `gamfit` doesn't work.
2. Creating the CDF with `makedist`.
3. Run the Kolmogorov-Smirnov test with `kstest`.

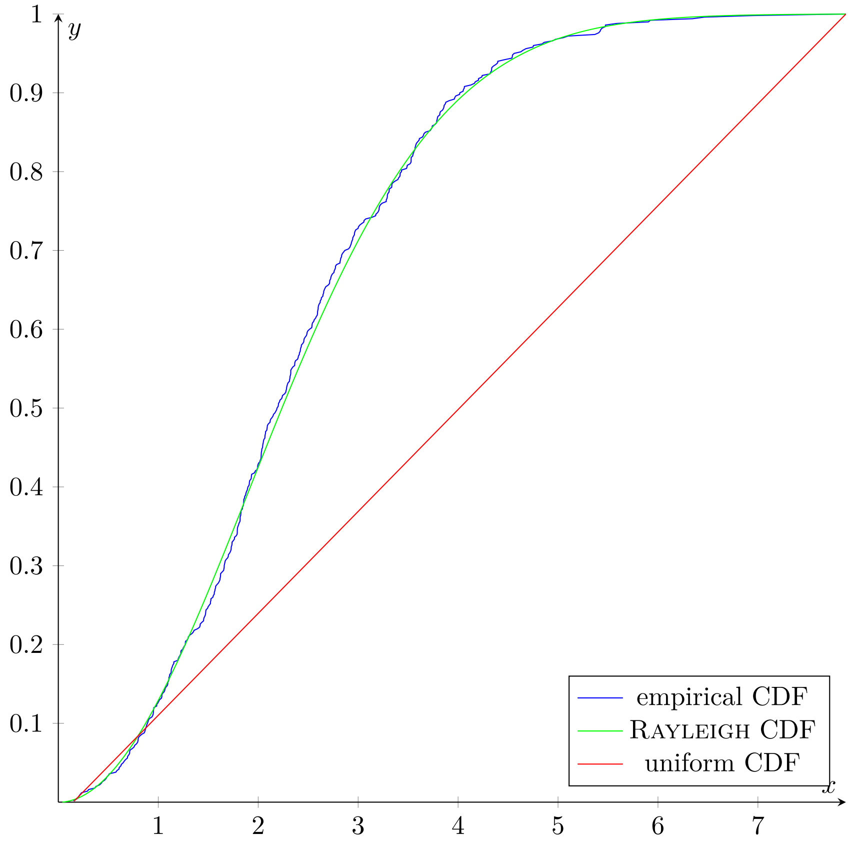
```
%normal distribution
[mu,sigma] = normfit(data)
norm_cdf = makedist('Normal','mu',mu,'sigma',sigma);
[h,p] = kstest(data,'CDF',norm_cdf)

%exponential distribution
mu = expfit(data)
exp_cdf = makedist('Exponential','mu',mu);
[h,p] = kstest(data,'CDF',exp_cdf)

%uniform distribution
[low,up] = unifit(data)
uni_cdf = makedist('Uniform','lower',low,'upper',up);
[h,p] = kstest(data,'CDF',uni_cdf)

%lognormal distribution
logmu = mean(log(data))
logsigma = std(log(data))
logn_cdf = makedist('Lognormal','mu',logmu,'sigma',logsigma);
[h,p] = kstest(data,'CDF',logn_cdf)

%rayleigh distribution
b = raylfit(data)
rayl_cdf = makedist('Rayleigh','b',b);
[h,p] = kstest(data,'CDF',rayl_cdf)

%gamma distribution
```

```
distribution = fitdist(data','Gamma');
a = distribution.a
b = distribution.b
gamma_cdf = makedist('Gamma','a',a,'b',b);
[h,p] = kstest(data,'CDF',gamma_cdf)
```

Running this gives the following output. The best fitting distribution is marked green, the worst red.

distribution	estimated parameters	Kolmogorov-Smirnov test	
		h	p
normal	$\mu = 2.3804$, $\sigma = 1.2486$	$h = 1$	$p = 0.0158$
exponential	$\mu = 2.3804$	$h = 1$	$p = 2.2618 \cdot 10^{-23}$
uniform	$\mathit{lower} = 0.1478$, $\mathit{upper} = 7.8807$	$h = 1$	$p = 1.5096 \cdot 10^{-72}$
lognormal	$\log(\mu) = 0.7050$, $\log(\sigma) = 0.6243$	$h = 1$	$p = 0.0017$
Rayleigh	$b = 1.9003$	$h = 0$	$p = 0.8939$
gamma	$a = 3.2378$, $b = 0.7352$	$h = 0$	$p = 0.2771$

3.2 Part (2)

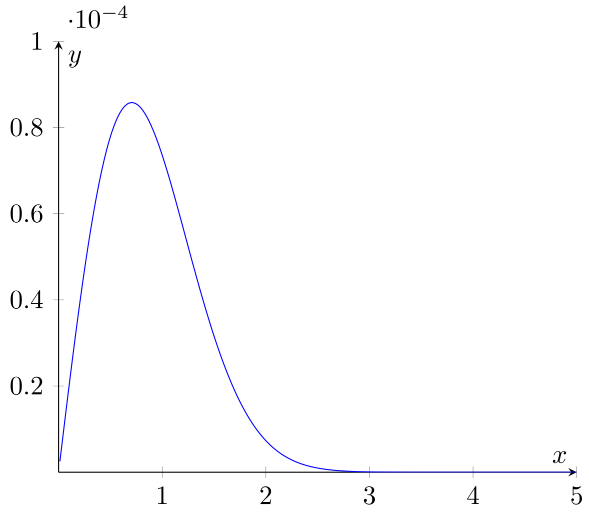


4 Task 4

4.1 Part (1)

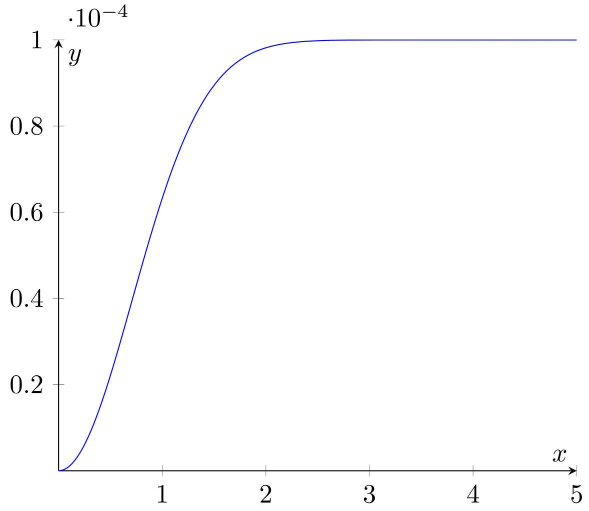
The probability density function $f(t)$ is

$$f(t) = \frac{2t \cdot \frac{\exp(-t^2)}{100}}{100} = \frac{t \cdot \exp(-t^2)}{5000}$$



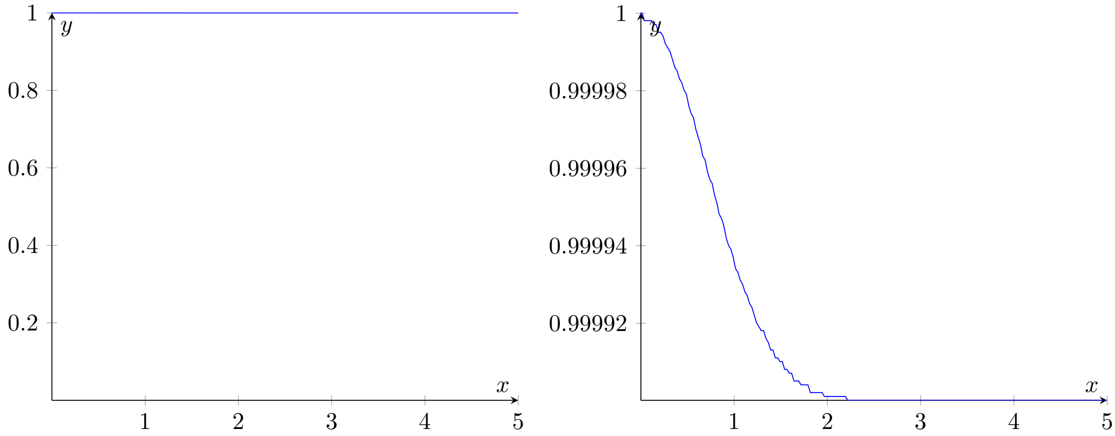
The cumulative distribution function $F(t)$ is then

$$\begin{aligned} F(t) &= \int_0^t f(\xi)\,\mathrm{d}\xi \\ &= \int_0^t \frac{\xi \cdot \exp(-\xi^2)}{5000}\,\mathrm{d}\xi \\ &= \frac{\exp(-t^2)\Big(\exp(t^2) - 1\Big)}{10000} \end{aligned}$$



For the survival function we get

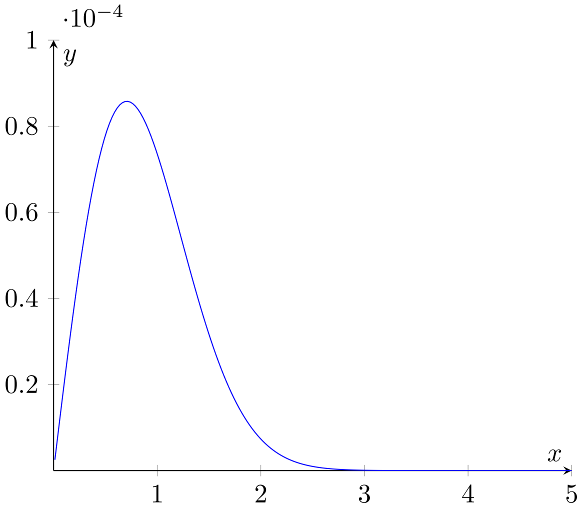
$$\begin{aligned} R(t) &= 1 - F(t) \\ &= \frac{\exp(-t^2) + 9999}{10000} \end{aligned}$$



To get the reliability of the component at $t = 7$ we simply evaluate $R(7)$ which is 0.9999.

The hazard function is defined as

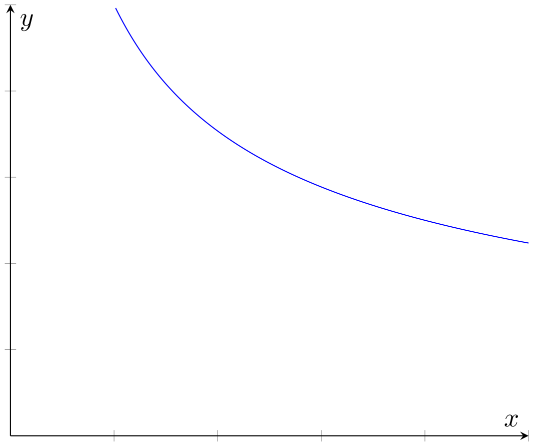
$$\begin{aligned} h(t) &= \frac{f(t)}{1 - F(t)} \\ &= \frac{2t}{9999 \cdot \exp(t^2) + 1} \end{aligned}$$



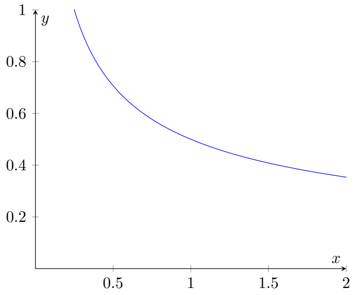
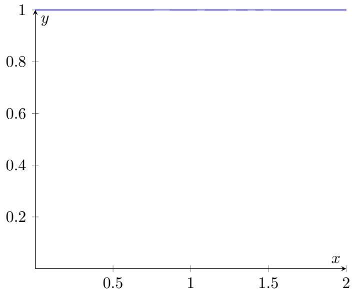
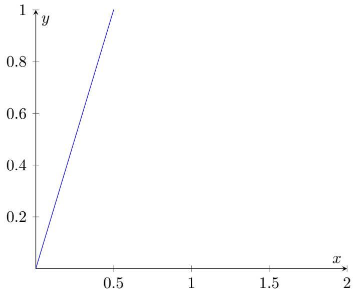
The hazard function describes how an item ages where t affects the risk of failure. It is the frequency with which the item fails, expressed in failures per unit of time.

4.2 Part (2)

Given $h(x) \sim (\sqrt{x})^{-1}$ we will try to find out the *shape*-parameter of the WEIBULL distribution first.



Comparing this graph to graphs of the hazard function with different *shape*-parameters we see that *shape* = 0.5 fits best.

$shape = 0.5$	$shape = 1$	$shape = 2$
Hazard function $\left(h = \frac{\text{PDF}}{1-\text{CDF}}\right)$		
		

To get the *scale*-parameter of the distribution we use the other provided information:

$$\begin{aligned} 5 &= \mu \\ &= scale \cdot \Gamma\left(1 + \frac{1}{shape}\right) \\ &= scale \cdot \Gamma(3) \\ \Rightarrow scale &= \frac{5}{2} \end{aligned}$$

Let's build the survival function:

$$\begin{aligned} R(t) &= 1 - \left(1 - \exp\left(-\sqrt{\frac{x}{5/2}}\right)\right) \\ &= \exp(-\sqrt{x} \cdot \sqrt{2.5}) \end{aligned}$$

That mean that the probability of surviving 6 years (30 years) is $R(6) = 0.0208$ ($R(30) = 0.0002$).