

# Applied statistics: Coursework 1

HENRY HAUSTEIN

1st March 2019

## Contents

<b>1</b>	<b>Task 1</b>	<b>1</b>
1.1	Part (1) . . . . .	1
1.2	Part (2) . . . . .	1
1.3	Part (3) . . . . .	1
1.4	Part (4) . . . . .	1
<b>2</b>	<b>Task 2</b>	<b>2</b>
2.1	Part (1) . . . . .	2
2.2	Part (2) . . . . .	2
2.3	Part (3) . . . . .	3
<b>3</b>	<b>Task 3</b>	<b>5</b>
3.1	Part (1) . . . . .	5
3.2	Part (2) . . . . .	6
<b>4</b>	<b>Task 4</b>	<b>7</b>
4.1	Part (1) . . . . .	7
4.2	Part (2) . . . . .	8

## 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:

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

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)

To get the 95% confidence interval via bootstrap I want to use the `bootci` function in MATLAB.

```
1 data = [3.75, 4.05, 3.81, 3.23, 3.13, 3.3, 3.21, 3.32, ...
2 4.09, 3.9, 5.06, 3.85, 3.88, 4.06, 4.56, 3.6, 3.27, ...
3 4.09, 3.38, 3.37, 2.73, 2.95, 2.25, 2.73, 2.55, 3.06];
4 parameter = @(y) length(find(y > 4.5))/length(y);
5
6 bootci(10000, {parameter, data}, 'alpha', 0.05, 'type', ...
7 'percentile')
```

That gives the 95% confidence interval: [0, 0.1923]

### 1.4 Part (4)

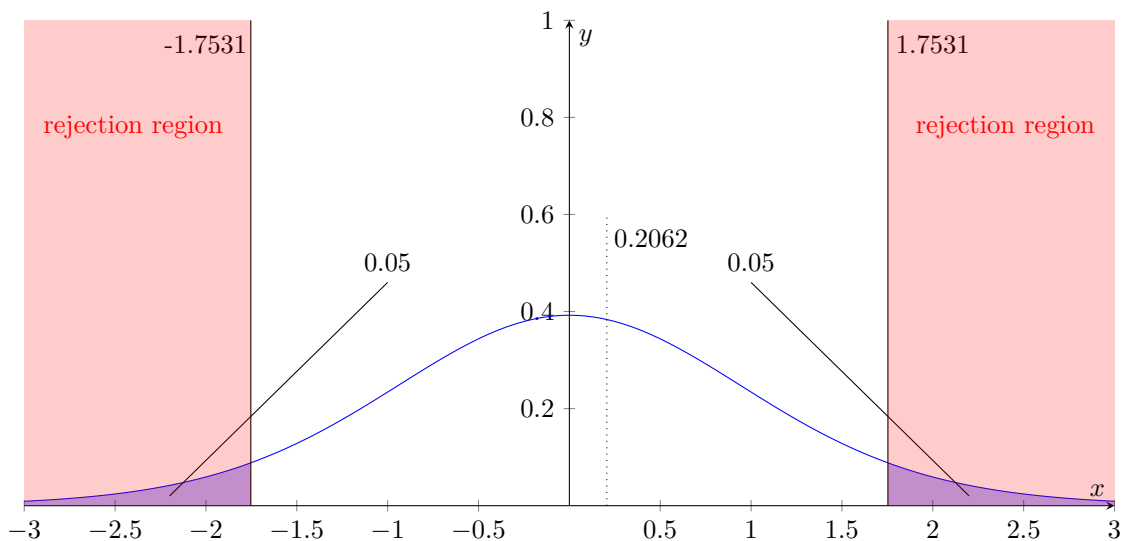
Yes, the confidence interval from the bootstrap procedure is more appropriate because it's not containing Al/Be ratios that are not possible like -0.0276. A negative ratio would suggest that there is a negative amount of data points in the sample which exceed 4.5. That is not possible.

## 2 Task 2

### 2.1 Part (1)

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

null hypothesis	$H_0: \mu = 0$
alternative hypothesis	$H_A: \mu \neq 0$
t-test for $\mu$	$t = \frac{m-0}{\frac{s}{\sqrt{15}}} = \frac{0.0853}{\frac{1.6031}{\sqrt{15}}} = 0.2062$
rejection region	$\text{tinv}(0.05, 15) = -1.7531$
conclusion	$t$ is 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  $\text{tinv}(0.025, 15) = -2.1314$ . 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  $\text{type 2 error} = 1 - \text{power}$ .

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

## 2. Task 2

---

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)

$H_0$ :  $\mu = 0$ , normal distribution, small model  $M_S$

$H_A$ :  $\mu \neq 0$ , normal distribution, big model  $M_B$

The log-likelihood function for normal distribution is

$$-\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2 \quad (1)$$

Let's start with the MLEs for  $\mu$  and  $\sigma$  in  $M_B$ :

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{j=1}^n x_j \\ &= 0.0853 \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{j=1}^n (x_j - \hat{\mu})^2 \\ &= 2.3986 \end{aligned}$$

Maximum possible value for the log-likelihood  $\xrightarrow{eq. (1)} -27.8457$ .

Now we'll calculate the MLE for  $\sigma$  in  $M_S$ :

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \sum_{j=1}^n (x_j - \hat{\mu})^2 \\ &= 2.3986 \end{aligned}$$

Maximum possible value for the log-likelihood  $\xrightarrow{eq. (1)} -27.8684$ .

Likelihood ratio test:

$$\begin{aligned} \chi^2 &= 2(l(M_B) - l(M_S)) \\ &= 0.0454 \end{aligned}$$

It should be compared to  $\chi^2(1 \text{ degree of freedom})$  since the difference in unknown parameters is equal to 1. The following piece of MATLAB code will calculate the p-value.

```
1 p = chi2cdf(0.0454, 1, 'upper')
```

The p-value is 0.8313 which means that we accept  $H_0$ : The small model  $M_S$  fits the data good enough. This is the same result as in part (1) and (2).

## 3 Task 3

### 3.1 Part (1)

First of all we need to prepare the data:

```
1 raw = load('input_data.txt');
2 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`.

```
1 %normal distribution
2 [mu,sigma] = normfit(data)
3 norm_cdf = makedist('Normal','mu',mu,'sigma',sigma);
4 [h,p] = kstest(data,'CDF',norm_cdf)
5
6 %exponential distribution
7 mu = expfit(data)
8 exp_cdf = makedist('Exponential','mu',mu);
9 [h,p] = kstest(data,'CDF',exp_cdf)
10
11 %uniform distribution
12 [low,up] = unifit(data)
13 uni_cdf = makedist('Uniform','lower',low,'upper',up);
14 [h,p] = kstest(data,'CDF',uni_cdf)
15
16 %lognormal distribution
17 logmu = mean(log(data))
18 logsigma = std(log(data))
19 logn_cdf = makedist('Lognormal','mu',logmu,'sigma',logsigma);
20 [h,p] = kstest(data,'CDF',logn_cdf)
21
22 %rayleigh distribution
23 b = raylfit(data)
24 rayl_cdf = makedist('Rayleigh','b',b);
25 [h,p] = kstest(data,'CDF',rayl_cdf)
26
27 %gamma distribution
```

### 3. Task 3

---

```

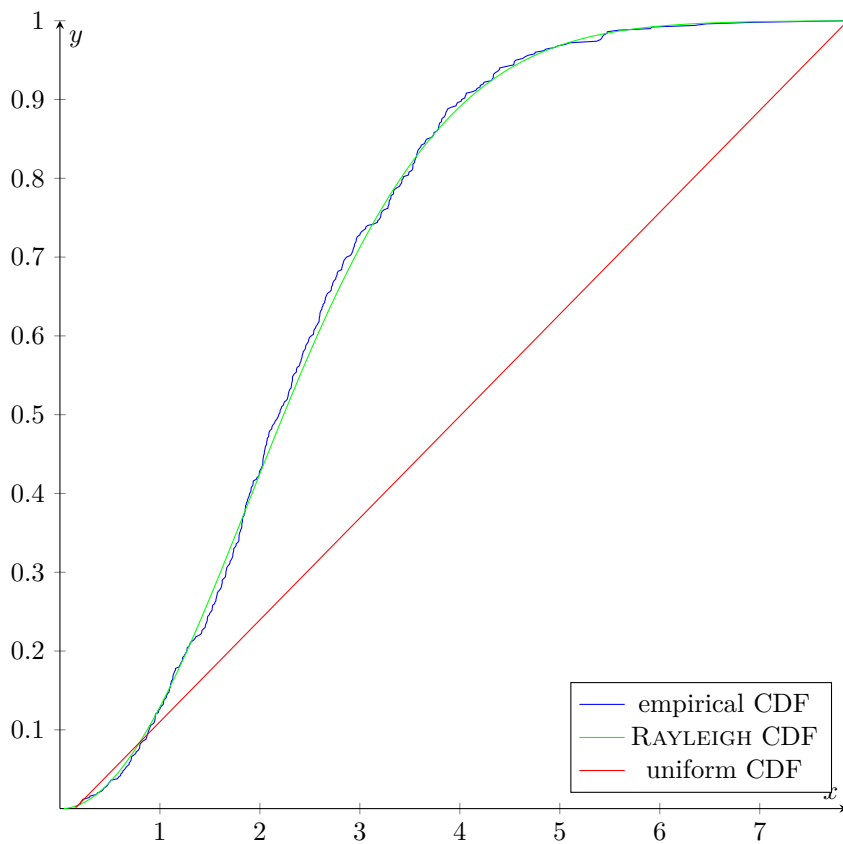
28 distribution = fitdist(data, 'Gamma');
29 a = distribution.a
30 b = distribution.b
31 gamma_cdf = makedist('Gamma','a',a,'b',b);
32 [h,p] = kstest(data,'CDF',gamma_cdf)

```

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	$lower = 0.1478, upper = 7.8807$	$h = 1$	$p = 1.5096 \cdot 10^{-72}$
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)

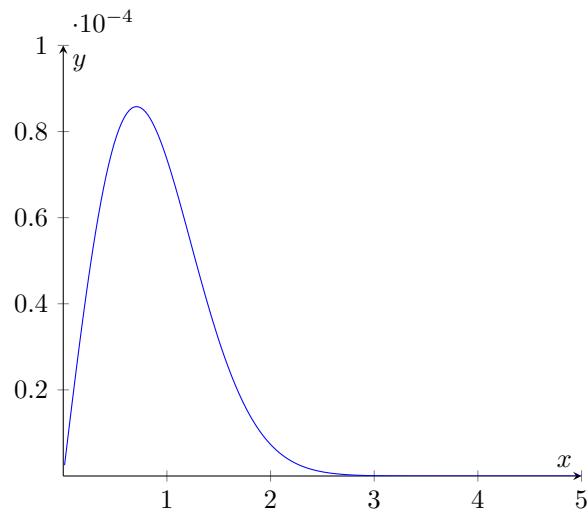


## 4 Task 4

### 4.1 Part (1)

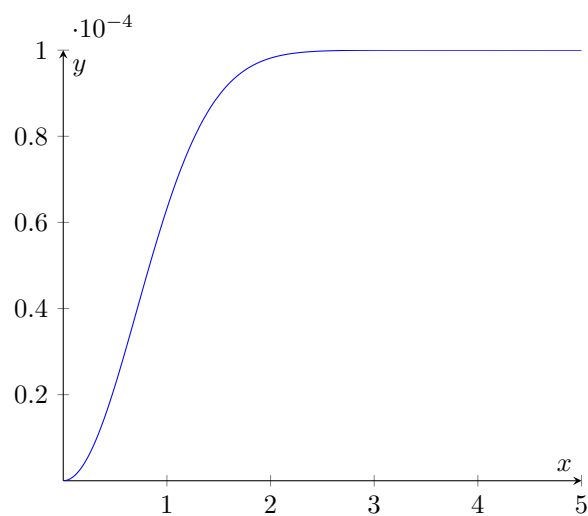
The probability density function  $f(t)$  is

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



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

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

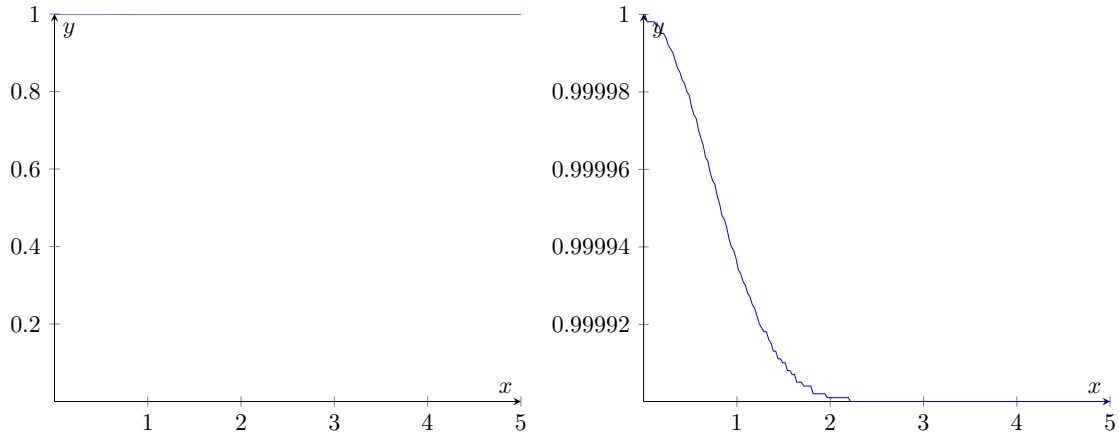


#### 4. Task 4

---

For the survival function we get

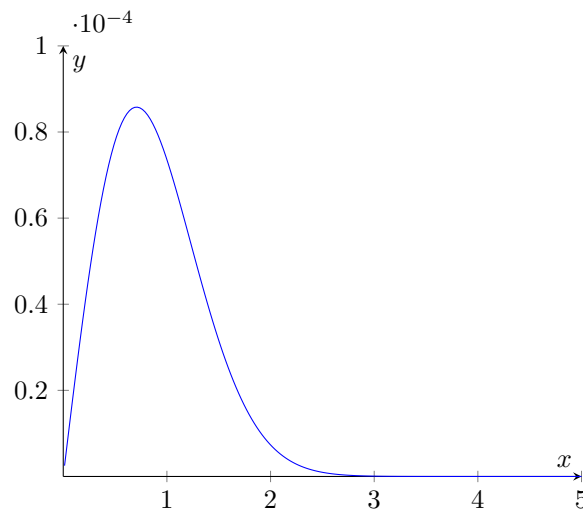
$$\begin{aligned} R(t) &= 1 - F(t) \\ &= \frac{\exp(-t^2) + 9999}{10000} \end{aligned}$$



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.

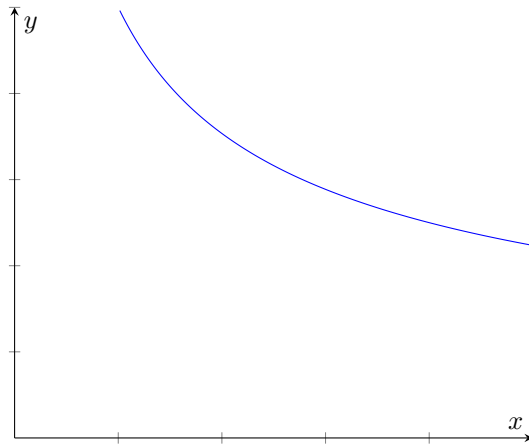
#### 4.2 Part (2)

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

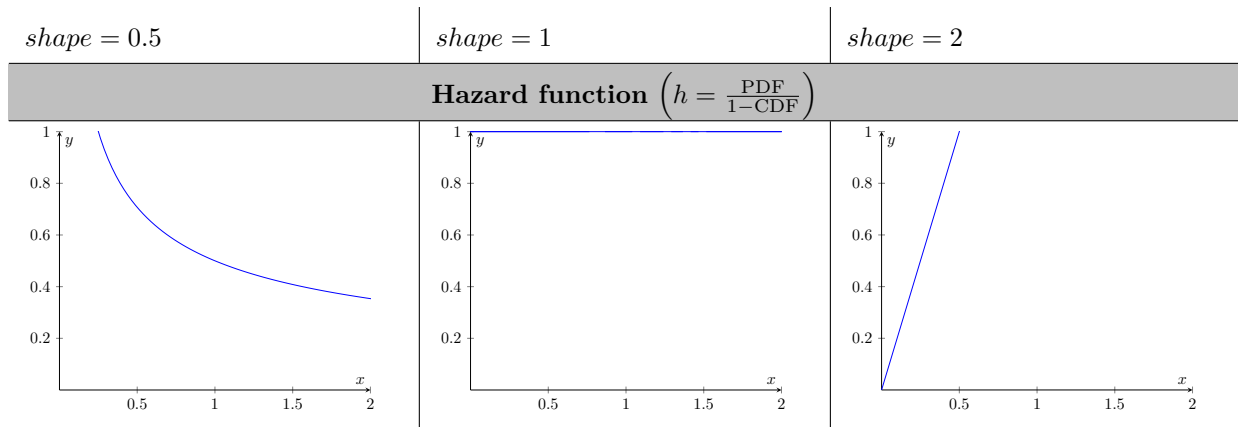


#### 4. Task 4

---



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



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  $R(6) = 0.0208$  ( $R(30) = 0.0002$ ).