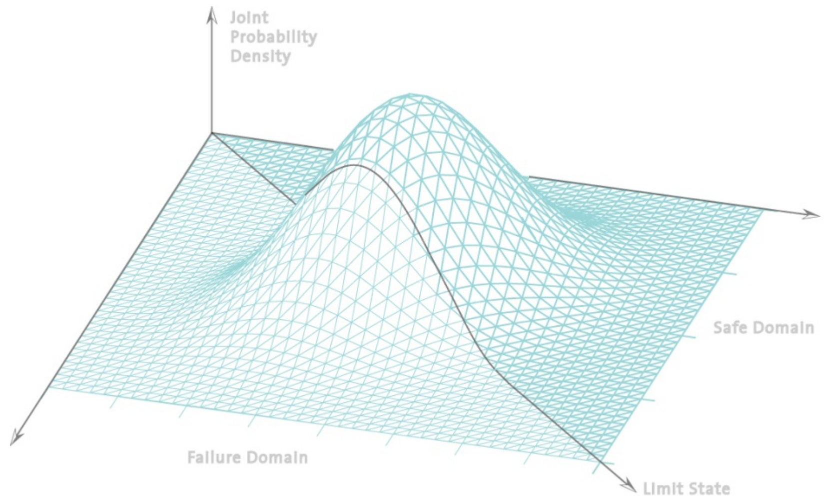


Exercise Tutorial 11

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber
Swiss Federal Institute of Technology Zurich
ETHZ

- Result of the 2nd assessment is announced on 25th June.
- Check of correction of the 2nd assessment is offered on 27th June 10 – 11AM, otherwise upon appointments.
- Complete solutions of the assessments are uploaded on 10th July.
- Revised script with the list of modifications is uploaded on 10th July.
- Final examination is held on 7th September.
- Office hours during semester vacation:

Exercise 9.3 (Group exercise)

It is known that the data shown in the table are the realizations of a random variable X characterized by the cumulative distribution function,

$F_X(x) = x^{\alpha}, 0 \leq x \leq 1$ with unknown parameter α.

Estimate the parameter α in the following methods.

a. Estimate the unknown parameter α with the *method of moments*.

b. Estimate the unknown parameter α with the *maximum likelihood method*.

c. Draw the cumulative distribution function with the estimated parameter in (b.) and the observed cumulative distribution.

No. of data	Realization
1	0.22
2	0.97
3	0.92
4	0.59
5	0.39
6	0.74
7	0.81
8	0.86
9	0.39
10	0.67

Exercise 9.3 (Group exercise)

a) Method of moments

“sample first moment = analytical first moment”

The sample first moment is calculated by

$$\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i = 0.656$$

The analytical first moment is calculated by

$$\int x f_X(x|\alpha)dx = \int_0^1 x\alpha x^{\alpha-1}dx = \left[\frac{\alpha}{\alpha+1}x^{\alpha+1}\right]_0^1 = \frac{\alpha}{\alpha+1}$$

Equate and solve

$$\frac{\alpha}{\alpha+1} = 0.656 \qquad \Leftrightarrow \alpha = 1.91$$

Exercise 9.3 (Group exercise)

b) Maximum likelihood method

The likelihood function is described as:

$$L(\alpha) = \prod_{i=1}^{n} f(\hat{x}_i \mid \alpha)$$

Maximize the log likelihood function with respect to α:

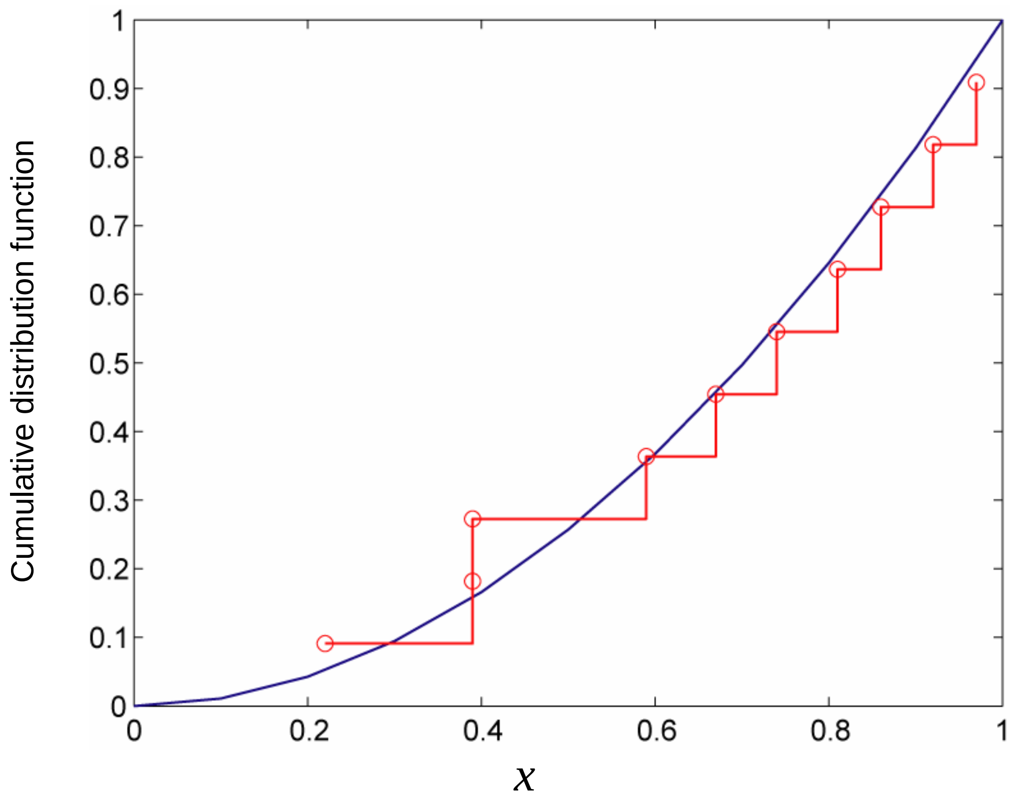
No. of data	Realization
1	0.22
2	0.97
3	0.92
4	0.59
5	0.39
6	0.74
7	0.81
8	0.86
9	0.39
10	0.67

$$l(\alpha) = \sum_{i=1}^{n} \ln(f(\hat{x}_i \mid \alpha)) = \sum_{i=1}^{n} \ln(\alpha \hat{x}_i^{\alpha-1}) = n \ln \alpha + (\alpha - 1) \sum_{i=1}^{n} \ln \hat{x}_i$$

$$\frac{\partial l(\alpha)}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n} \ln \hat{x}_i = 0 \Leftrightarrow \alpha = -\frac{n}{\sum_{i=1}^{n} \ln \hat{x}_i} = 1.96$$

Exercise 9.3 (Group exercise)

c)

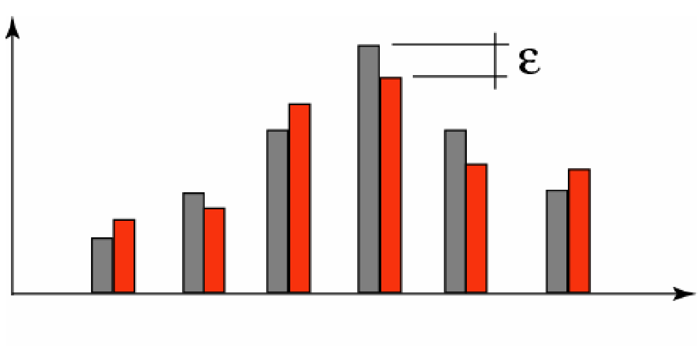


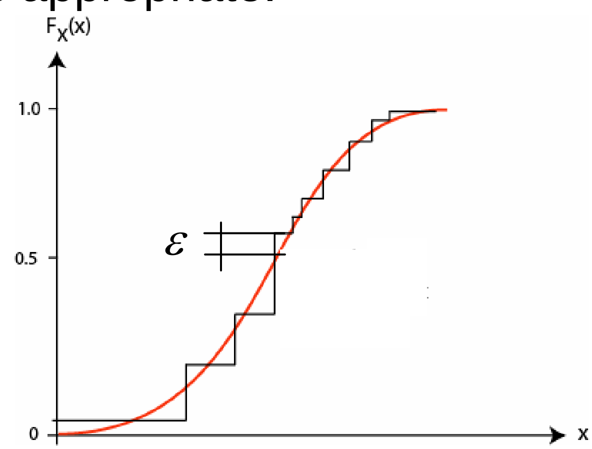
The essence of the goodness-of-fit test is to
measure the deviation between assumed distribution and observations.

There are many ways to measure the deviation.

Let K be a random variable representing the deviation. We judge with K whether the assumed distribution model the observations well or not.

The operating rule is then simply,
Accept H_0 if $K<c$
Reject H_0 otherwise,
where H_0 is that the assumed distribution is appropriate.





Goodness-of-fit test

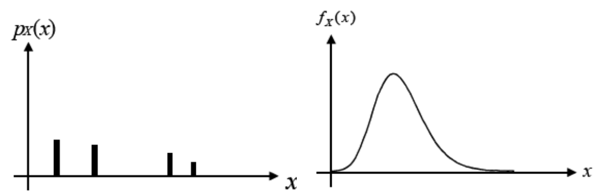
What can be a good measure of the deviation K?

The choice of K can be almost arbitrary as long as it is related to the deviation, but it must be the one whose distribution is known, otherwise we can not determine c in accordance with the level of significance α.

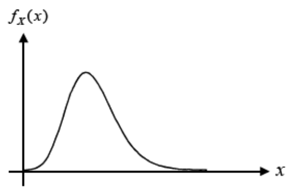
$$P[K > c \mid \theta \in \Theta_0] = \alpha \Leftrightarrow P[K \le c \mid \theta \in \Theta_0] = 1 - \alpha$$

We consider two alternatives:

1) Chi-square test (for any cases*)
*We discretize in continuous cases.



2) Kolmogorov-Smirnov test (only for continuous case)



Let's begin with a simple exercise.

Exercise 10.1

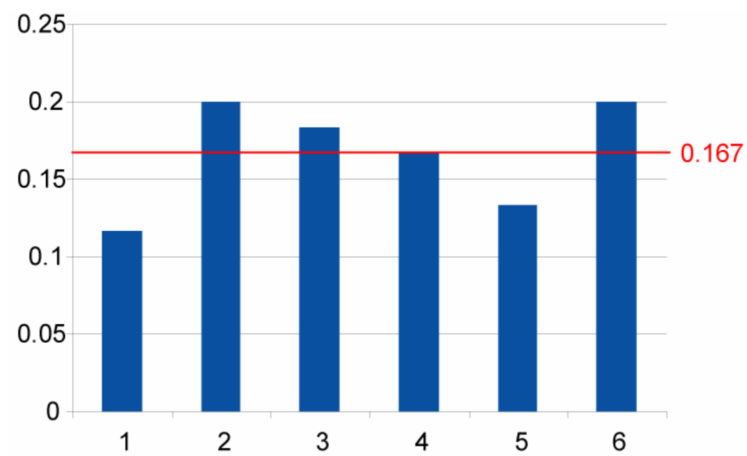
A dice is suspected to be asymmetric, resulting in the inhomogeneity of probability that each side of a dice comes out. In order to judge this suspicion statistically, 60 trials were made and the result is shown in the table.

1) Draw the relative frequency histogram, and compare with the uniform probability density function under the assumption that the dice is symmetric.
2) What is the probability that each side of a dice comes out 10 times respectively in 60 trials when the dice is symmetric?
3) Set the symmetry property of the dice as the null hypothesis, and test the hypothesis with the χ^2 test at the 5% level of significance.

Side	No. of realizations
1	7
2	12
3	11
4	10
5	8
6	12
Sum	60

1) Comparison between the observation and the assumed distribution.

Spot of dice	No. of realization	Relative frequency
1	7	0.1167
2	12	0.2
3	11	0.1833
4	10	0.1667
5	8	0.1333
6	12	0.2
Sum	60	1



There is deviation…but is this too large to say that the dice is asymmetric?

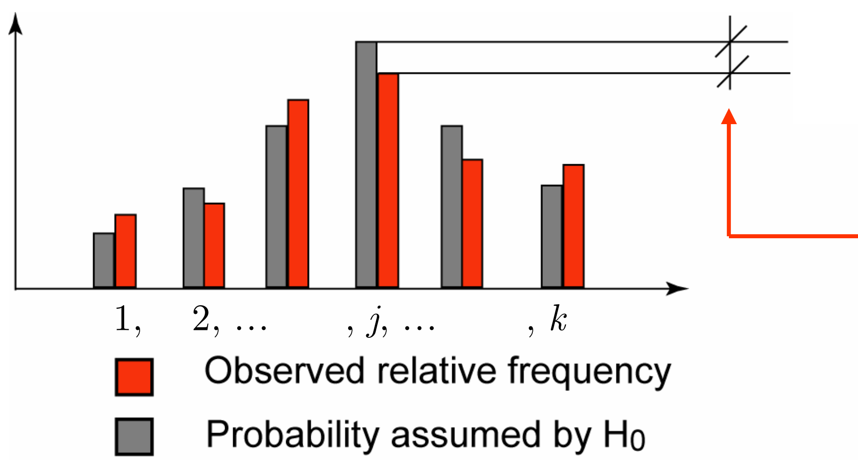
First of all, calculate the probability that the numbers of outcomes completely coincide the expected numbers under the assumption that the dice is symmetric.

The numbers that each side comes out follow the polynomial distribution:

$$
\begin{aligned}
P[N_{o,j} &= 10,\ j = 1,2,3,4,5,6] \\
&= \frac{60!}{10!10!10!10!10!10!}\left(\frac{1}{6}\right)^{10}\left(\frac{1}{6}\right)^{10}\left(\frac{1}{6}\right)^{10}\left(\frac{1}{6}\right)^{10}\left(\frac{1}{6}\right)^{10}\left(\frac{1}{6}\right)^{10} \\
&= \frac{60!}{(10!)^6}\left(\frac{1}{6}\right)^{60} \\
&= 0.0000745
\end{aligned}
$$

The probability is very very small!!! We should allow the deviation to some extent.

Chi-square test



$$K=\sum_{j=1}^{k}\overset{\text{Weight}}{c_j}\underbrace{\left(p(x_j)-\frac{N_{o,j}}{n}\right)^2}_{\text{Deviation}}$$

Fundamentally, the weights c_j are arbitrary but if you choose $c_j=\dfrac{n}{p(x_j)}$

then it is known that K follows approximately χ^2 distribution with k - 1 degrees of freedom.

When you know the distribution of K, you can select c in accordance with the level of significance. ($P[K>c\,|\,\theta\in\Theta_0]=\alpha \Leftrightarrow P[K\le c\,|\,\theta\in\Theta_0]=1-\alpha$)

We adopt $K = \varepsilon_m^{\ 2}$ as the sample statistic.

$$\varepsilon_m^{\ 2} = \sum_{j=1}^{k} \frac{n}{p(x_j)} \left(p(x_j) - \frac{N_{o,j}}{n} \right)^2 = \sum_{j=1}^{k} \frac{\left(np(x_j) - N_{o,j} \right)^2}{np(x_j)} = \sum_{j=1}^{k} \frac{\left(N_{p,j} - N_{o,j} \right)^2}{N_{p,j}} \quad \text{(Script E.67)}$$

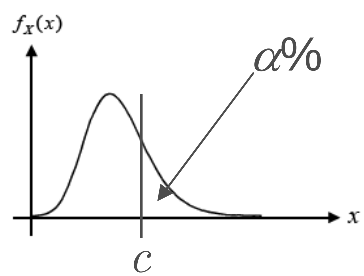
Again $\varepsilon_m^{\ 2}$ follows χ^2 distribution with $k-1$ degrees of freedom.

Therefore it is possible to determine c in accordance with the level of significance.

$$P[\varepsilon_m^{\ 2} > c \mid \theta \in \Theta_0] = \alpha \Leftrightarrow P[\varepsilon_m^{\ 2} \le c \mid \theta \in \Theta_0] = 1 - \alpha$$

In the case of this example, k = 6 and α = 5%.

f	$\chi^2_{F=0.01}$	$\chi^2_{F=0.05}$	$\chi^2_{F=0.10}$	$\chi^2_{F=0.25}$	$\chi^2_{F=0.50}$	$\chi^2_{F=0.75}$	$\chi^2_{F=0.90}$ α=10%	$\chi^2_{F=0.95}$ α=5%
1	0.0002	0.0039	0.0158	0.1015	0.4549	1.3233	2.7055	3.8415
2	0.0201	0.1026	0.2107	0.5754	1.3863	2.7726	4.6052	5.9915
3	0.1148	0.3518	0.5844	1.2125	2.3660	4.1083	6.2514	7.8147
4	0.2971	0.7107	1.0636	1.9226	3.3567	5.3853	7.7794	9.4877
5	0.5543	1.1455	1.6103	2.6746	4.3515	6.6257	9.2363	11.0705
6	0.8721	1.6354	2.2041	3.4546	5.3481	7.8408	10.6446	12.5916



H$_0$: $p_j = 1/6$ (j=1,2,3,4,5,6)
H$_1$: $p_j \neq 1/6$ at least one of j = 1,2,3,4,5,6.

Calculate the value of the sample statistic $\varepsilon_m^{\ 2}$ based on 60 observations.

Side	$N_{o,j}$	$N_{p,j} = np(x_j)$	$(N_{j,o} - N_{p,j})^2$	ε_m^2
1	7	10	9	9/10
2	12	10	4	4/10
3	11	10	1	1/10
4	10	10	0	0/10
5	8	10	4	4/10
6	12	10	4	4/10
Sum	60			$= 2.20$

$$\varepsilon_m^{\ 2} = 2.2 < 11.07 = c$$

H$_0$ is not rejected, meaning that the dice may be symmetric.

Exercise 10.2

For the estimation of the concrete compressive strength of a certain concrete production, 20 samples were measured and the result is shown in the table. It is assumed that the concrete compressive strength follows the Normal distribution.

a. Estimate the unknown parameters of the distribution with the method of moments.

b. Test the goodness of fit for the distribution with estimated parameters with the test at the 5% significance level.

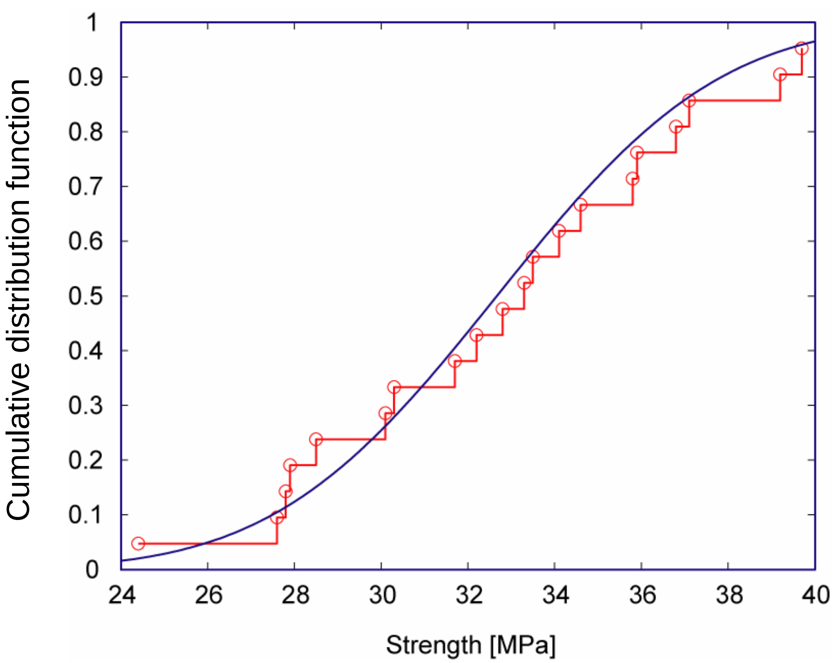
No. of sample (i)	Compressive strength (MPa)	No. of sample (i)	Compressive strength (MPa)
1	24.4	11	33.3
2	27.6	12	33.5
3	27.8	13	34.1
4	27.9	14	34.6
5	28.5	15	35.8
6	30.1	16	35.9
7	30.3	17	36.8
8	31.7	18	37.1
9	32.2	19	39.2
10	32.8	20	39.7

Exercise 10.2

a. Estimate the unknown parameters of the distribution with the method of moments.

$$\mu = m_1 = 32.67$$

$$\sigma = \sqrt{m_2 - m_1^2} = \sqrt{1083.4 - 32.67^2} = 4.04$$



Exercise 10.2

b. Test the goodness of fit for the distribution with estimated parameters with the test at the 5% significance level.

When the χ^2-test is applied for continuous cases, we have to disretize the intervals.

$$\varepsilon_m^2 = \sum_{j=1}^{k} \frac{(N_{o,j} - N_{p,j})^2}{N_{p,j}}$$

Interval	$N_{o,j}$	$p(x_j)$	$N_{p,j} = np(x_j)$	ε_m^2
-30				
30-33				
33-36				
36-				
Sum				

Exercise 10.2

b. Test the goodness of fit for the distribution with estimated parameters with the test at the 5% significance level.

Interval	$N_{o,j}$	$p(x_j)$	$N_{p,j} = np(x_j)$	ε_m^2
-30	5	$\Phi\left(\frac{30-32.67}{4.04}\right) = 0.254$	5.08	0.001
30-33	5	$\Phi\left(\frac{33-32.67}{4.04}\right) - \Phi\left(\frac{30-32.67}{4.04}\right) = 0.278$	5.56	0.06
33-36	6	$\Phi\left(\frac{36-32.67}{4.04}\right) - \Phi\left(\frac{33-32.67}{4.04}\right) = 0.263$	5.26	0.10
36-	4	$1-\Phi\left(\frac{36-32.67}{4.04}\right) = 0.205$	4.10	0.002
Sum	20			= 0.163

Exercise 10.2

b. Test the goodness of fit for the distribution with estimated parameters with the test at the 5% significance level.

The sample statistic follows the Chi-square distribution with 4-1-2=1 degree of freedom. At the 5% significance level, the null hypothesis should be rejected if the sample statistic is larger than 3.84, see the probability table for the Chi-square distribution (Annex T, Table T.3).

Since the sample statistic is obtained as 0.163 from the observations the null hypothesis cannot be rejected at the 5% significance level, i.e., the assumption that the concrete compressive strength may follow the Normal distribution with the mean 32.67MPa and the standard deviation 4.04MPa cannot be rejected.

Exercise 10.4

The strength of 30 wood samples has been measured and the results are shown in the table. The strength is assumed to follow an Exponential distribution.

a. Estimate the parameter of the Exponential distribution using the method of moments.
b. Draw the cumulative distribution function with the estimated parameters, together with the observed cumulative distribution.
c. Test the goodness of fit for the Exponential distribution with the χ^2 test at the 10% significance level.
d. Test the goodness of fit for the Exponential distribution with the Kolmogorov-Smirnov test at the 10% significance level. The parameter of the Exponential distribution is assumed to be equal to λ=0.04.

Exercise 10.4

The strength of 30 wood samples has been measured and the results are shown in the table. The strength is assumed to follow an Exponential distribution.

No.	Strength (MPa)	No.	Strength (MPa)	No.	Strength (MPa)
1	12.8	11	23.4	21	29.3
2	16.3	12	26.8	22	29.5
3	16.6	13	26.9	23	30.3
4	16.9	14	27	24	32.1
5	17.2	15	27.1	25	32.3
6	17.9	16	27.2	26	33.5
7	19.5	17	27.2	27	33.9
8	21.9	18	27.5	28	35.6
9	22.3	19	27.9	29	39.2
10	22.5	20	28.3	30	43.5

Exercise 10.4

a. Estimate the parameters of the exponential distribution using the method of moments.

The cumulative distribution function of the Exponential distribution is written as:

$$F_X(x) = 1 - \exp(-\lambda x)$$

The first moment is obtained as:

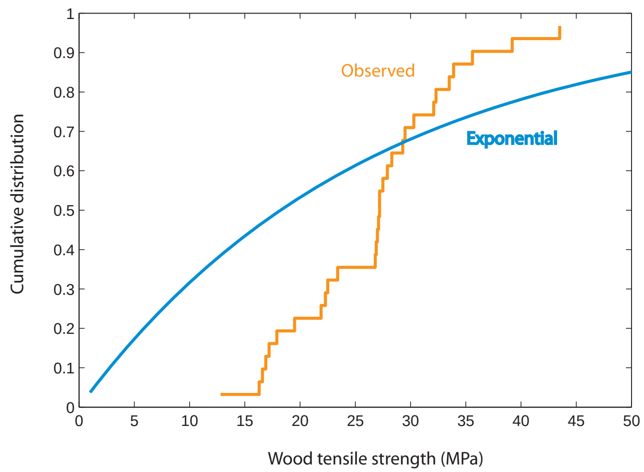
$$m_1 = \frac{1}{\lambda}$$

The parameter is estimated as:

$$\lambda = \frac{1}{m_1} = \frac{1}{1/30 \sum_{i=1}^{n} \hat{x}_i} = 0.038$$

Exercise 10.4

b. Draw the cumulative distribution function with the estimated parameter, together with the observed cumulative distribution.



Exercise 10.4

c. Test the goodness of fit for the Exponential distribution with the χ^2 test at the 10% significance level.

$$\varepsilon_m^2 = \sum_{j=1}^{k} \frac{(N_{o,j} - N_{p,j})^2}{N_{p,j}}$$

Interval	$N_{o,j}$	$p(x_j)$	$N_{p,j} = np(x_j)$	ε_m^2
-20	7	0.49	14.7	4.03
20-25	4	0.08	2.4	1.07
25-30	11	0.07	2.1	37.72
30-	8	0.36	10.8	0.73
Sum	30			43.55

Exercise 10.4

c. Test the goodness of fit for the Exponential distribution with the χ^2 test at the 10% significance level.

The sample statistic follows the Chi-square distribution with 4-1-1=2 degrees of freedom. At the 10% significance level, the null hypothesis shall be rejected if the sample statistic is larger than 4.6, see the probability table for the Chi-square distribution (Annex T, Table T.3).

Since the sample statistic is obtained as 43.55 from the observations, the null hypothesis should be rejected at the 10% significance level.

Exercise 10.4

d. Test the goodness of fit for the Exponential distribution with the Kolmogorov-Smirnov test at the 10% significance level. The parameter of the Exponential distribution is assumed to be equal to λ=0.04.

Kolmogorov-Smirnov test cannot be used when you estimate the parameters of distribution!!!

Exercise 10.4

d. Test the goodness of fit for the exponential distribution with the Kolmogorov-Smirnov test at the 10% significance level. The parameter of the Exponential distribution is assumed to be equal to λ=0.04.

i	x_i	$F_o(x_i^o)=\frac{i}{n}$	$F_p(x_i^o)$	$\lvert F_o(x_i^o)-F_p(x_i^o)\rvert$
1.0	12.8	0.033	0.401	0.367
2.0	16.3	0.067	0.479	**0.412**
3.0	16.6	0.100	0.485	0.385
4.0	16.9	0.133	0.491	0.358
5.0	17.2	0.167	0.497	0.331
6.0	17.9	0.200	0.511	0.311
7.0	19.5	0.233	0.542	0.308
8.0	21.9	0.267	0.584	0.317
9.0	22.3	0.300	0.590	0.290
10.0	22.5	0.333	0.593	0.260
11.0	23.4	0.367	0.608	0.241
12.0	26.8	0.400	0.658	0.258
13.0	26.9	0.433	0.659	0.226
14.0	27.0	0.467	0.660	0.194
15.0	27.1	0.500	0.662	0.162

	x_i	$F_o(x_i^o)=\frac{i}{n}$	$F_p(x_i^o)$	$\lvert F_o(x_i^o)-F_p(x_i^o)\rvert$
16.0	27.2	0.533	0.663	0.130
17.0	27.2	0.567	0.663	0.096
18.0	27.5	0.600	0.667	0.067
19.0	27.9	0.633	0.672	0.039
20.0	28.3	0.667	0.678	0.011
21.0	29.3	0.700	0.690	0.010
22.0	29.5	0.733	0.693	0.041
23.0	30.3	0.767	0.702	0.064
24.0	32.1	0.800	0.723	0.077
25.0	32.3	0.833	0.725	0.108
26.0	33.5	0.867	0.738	0.129
27.0	33.9	0.900	0.742	0.158
28.0	35.6	0.933	0.759	0.174
29.0	39.2	0.967	0.792	0.175
30.0	43.5	1.000	0.824	0.176

Exercise 10.4

d. Test the goodness of fit for the exponential distribution with the Kolmogorov-Smirnov test at the 10% significance level. The parameter of the Exponential distribution is assumed to be equal to λ=0.04.

At the 10% significance level and n=30, the null hypothesis should be rejected if the sample statistic is larger than 0.22, (Annex T, Table T.4).

Since the sample statistic is obtained as 0.412 from the observations the null hypothesis should be rejected at the 10% significance level.