Chapter (4): Model selection and evaluation.

Problem: data: $x_1, x_2, \ldots, x_n$.

the goal is to find an adequate distribution $F(x)$ to this data.

after doing some steps of drawing histograms, we decide to take a certain distribution $F^*(x)$.

Now we proceed using hypothesis testing if $F(x) \equiv F^*(x)$ or not.

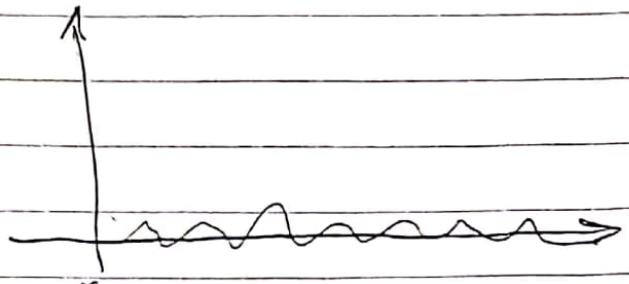
(1) Graphical methods:

* D(x)-plot:

we take the empirical distribution function $F_n(x)$.

$$F_n(x) = \frac{\sum_{i=1}^{n} 1_{(x_i \le x)}}{n} \xrightarrow{n \to \infty} F(x)$$

we define $D(x) = F_n(x) - F^*(x)$.



$F^*(x)$ is adequate.

$F^*(x)$ is not adequate.

Ex: 2, 3, 3, 3, 5, 8, 10, 13, 16.

An exponential distribution is fit to the data. We estimate the parameter using the MLE method.

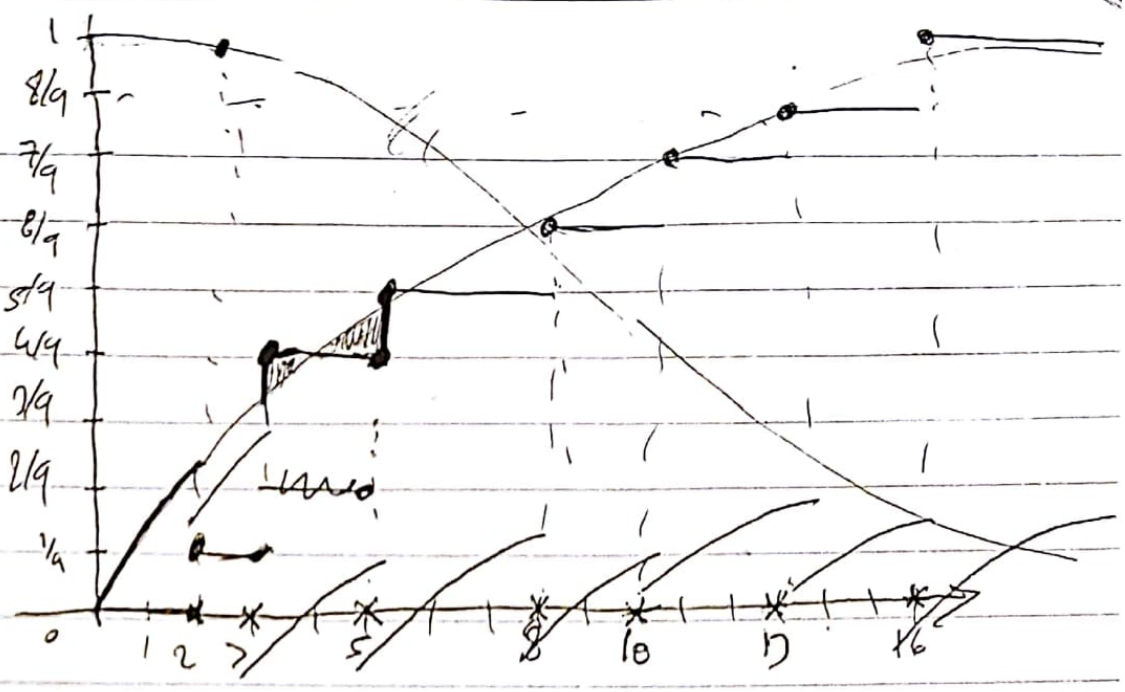
Plot $D(x)$?

$$\hat{\lambda} = \frac{1}{\bar{x}} = \frac{9}{63} = \frac{1}{7}$$

$$F^*(x) = 1 - e^{-x/7}, \quad x \geq 0.$$

$n=9$:

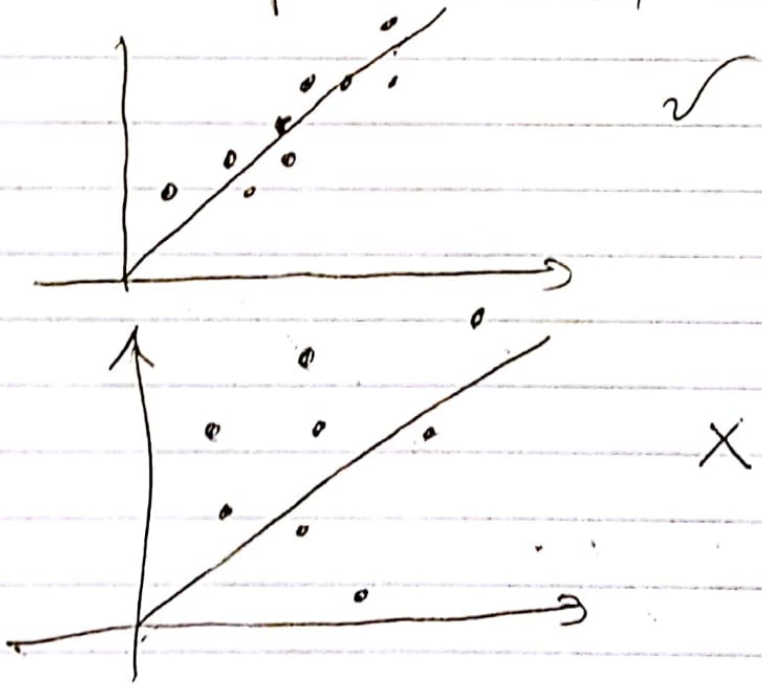
$$F_9(x) = \begin{cases} 0 & x < 2 \\ 1/9 & 2 \leq x < 3 \\ 4/9 & 3 \leq x < 5 \\ 5/9 & 5 \leq x < 8 \\ 6/9 & 8 \leq x < 10 \\ 7/9 & 10 \leq x < 13 \\ 8/9 & 13 \leq x < 16 \\ 1 & x \geq 16 \end{cases}$$



* P-P plot :

$x_1 < x_2 < \dots < x_n \rightarrow$ $\boxed{F_n(x_j) = \frac{j}{n+1}}$

we draw the points $(F_n(x_i), F^*(x_i))$.



✓

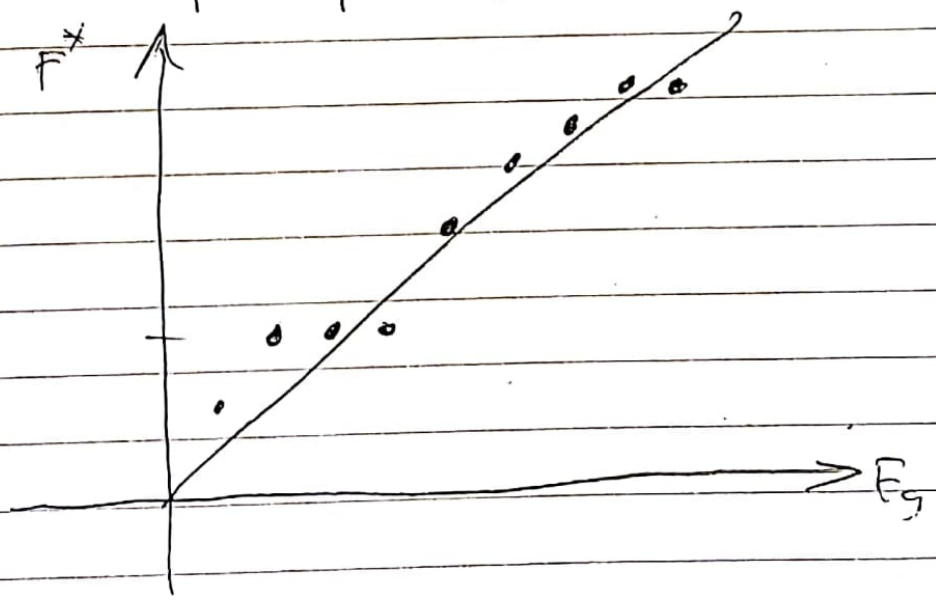
X

Ex: The same data.

~~Note~~ Draw the p-p plot.

$$x_1, \dots, x_n \Rightarrow x_{(1)}, \dots, x_{(n)}$$

j	x_j	$F_0(x_j)$	$F^*(x_j)$
1	2	0.1	0.249
2	3	0.2	0.349
3	3	0.3	0.349
4	3	0.4	0.349
5	5	0.5	0.510
6	8	0.6	0.681
7	10	0.7	0.760
8	12	0.8	0.844
9	16	0.9	0.883



(2) Hypothesis testing :

H_0 : null hypothesis.
H_a : alternative hypothesis.

	Accept H_0	Reject H_0.
H_0 is true	✓	type I error $= \alpha$
H_a is true	type II error	✓

α = significance level.

In general, one chooses a statistic T and accept H_0 if $T \leq c$,

$$P(T \leq c) = 1 - \alpha.$$

(3) Kolmogorov-Smirnov Hypothesis test (K-S test)

$H_0 : F \equiv F^*$; $H_a : F \neq F^*$.

statistic : $D = \sup_{x_1 \leq x \leq x_n} |F_n(x) - F^*(x)|$

$$D = \sup_{x_i} \left(|F_n(x_i) - F^*(x_i)| ; |F_n(x_{i-1}) - F^*(x_i)| \right).$$

level of significance (Confidence) α 0.1 0.05 0.01

Critical value $\frac{1.22}{\sqrt{n}}$ $\frac{1.36}{\sqrt{n}}$ $\frac{1.63}{\sqrt{n}}$.

6400

Ex: * 200, 400, 1000, 3000, 5000, 5400, 6200,

* $F^*(x) = 1 - e^{-\lambda x}$; $\hat{\lambda} = \frac{1}{\bar{x}}$ =

$\lambda = 1/3450$.

Determine whether the K-S test will result in rejection of null hypothesis with $\alpha = 10\%$.

j	x_j	$F_8(x_j) = \frac{j}{8}$	$F_8(x_{j-1})$	$F^*(x_j)$	Maximum
1	200	0.125	0	0.056	0.069
2	400	0.25	0.125	0.109	0.141
3	1000	0.375	0.25	0.252	0.123
4	3000	0.5	0.375	0.58	0.205
5	5000	0.625	0.5	0.765	0.265
6	5400	0.75	0.625	0.79	0.165
7	6200	0.875	0.75	0.834	0.084
8	6400	1	0.875	0.843	0.157

$D = 0.265$. ; $\frac{1.22}{\sqrt{8}} = 0.431$.

The acceptance of hypothesis.

(4) Anderson-Darling test:

$x_1, x_2, \ldots, x_n$.

$y_0 \le y_1 < y_2 < \cdots < y_k \le y_{k+1}$.

$$y_0 = \begin{cases} 0 & \text{if there is no truncation.} \\ d & \text{if there is.} \end{cases}$$

$$y_{k+1} = \begin{cases} \infty & \text{if there is no censoring.} \\ u & \text{if there is.} \end{cases}$$

The Anderson darling test statistic:

$$A^2 = -n F^*(u) + n\left[\sum_{j=0}^{k} \left(1 - F_n(y_j)\right)^2 \ln\left(\frac{1-F^*(y_j)}{1-F^*(y_{j+1})}\right)\right.$$

$$\left. + \sum_{j=1}^{k} \left(F_n(y_j)\right)^2 \ln\left(\frac{F^*(y_{j+1})}{F^*(y_j)}\right)\right].$$

level of Confidence α	0.10	0.05	0.01
Critical value	1.933	2.492	3.857

Ex: claim amount: 200; 400; 900; 1,600; 3,000; 5,000; 5,400; 6200.

$F^*(x) = 1 - e^{-\theta x}$; $1/\theta = 3,300$.

a) Find A^2.

b) Determine the result of the test for $\alpha = 10\%$.

j	y_j	$F_n(y_j)$	$F^*(y_j)$	a_j	b_j
0	0	0	0	0.06	0
1	200	0.125	0.058	0.046	0.010
2	400	0.25	0.114	0.102	0.051
3	1,000	0.375	0.261	0.071	0.054
4	1,600	0.5	0.394	0.106	0.160
5	3,000	0.625	0.597	0.085	0.104
6	5,000	0.75	0.780	0.007	0.017
7	5,400	0.875	0.805	0.0078	0.078
8	6,200	1	0.8472	0	0.1658
9	∞	1	1	0.482	0.553

$$a_j = (1 - F_n(y_j))^2 \ln\left(\frac{1 - F^*(y_j)}{1 - F^*(y_{j+1})}\right)$$

$$b_j = (F_n(y_j))^2 \ln\left(\frac{F^*(y_{j+1})}{F^*(y_j)}\right).$$

$$A^2 = -8 + 8(0.482 + 0.553) = 0.289$$

b) $c = 1.933$.

We accept the hypothesis.

(5) Chi-square test:

interval	# of observations
(c_0, c_1)	n_1
(c_1, c_2)	n_2
⋮	⋮
(c_{k-1}, c_k)	n_k

$$\hat{p}_j = F^*(c_j) - F^*(c_{j-1}).$$

$$E_j = n\hat{p}_j.$$

$$O_j = n_j.$$

The chi-square statistic:

$$\chi^2 = \sum_{j=1}^{k} \frac{(E_j - O_j)^2}{E_j} = \left(\sum_{j=1}^{k} \frac{n_j^2}{E_j} \right) - n.$$

χ^2 is computed from the chi-square distribution table with degree of freedom equal to $k-r-1$ where r is the number of parameters estimated in the model.

We accept the hypothesis if

$$\chi^2 \leq \chi^2_{k-r-1,\,1-\alpha}$$, α = level of Confidence

Ex. F:

interval	F(Cj)	# of observations
$x < 2$	0.035	5
$2 \leq x \leq 5$	0.170	42
$5 \leq x < 7$	0.630	137
$7 \leq x < 9$	0.83	66
$9 \leq x$	1	50
		n = 300.

a) χ^2.

b) Test the null hypothesis with $\alpha = 5\%$.

c) ——————————— $\alpha = 2.5\%$.

a)

j	$\hat{p}_j$	$E_j = n\hat{p}_j$	θ_j	$(E_j - \theta_j)^2$
1	0.035	10.5	5	30.25
2	0.095	28.5	42	182.25
3	0.5	150	137	169
4	0.2	60	66	36
5	0.17	51	50	1

$$\chi^2 = \frac{30.25}{10.5} + \frac{182.25}{28.5} + \frac{169}{150} + \frac{36}{60} + \frac{1}{51}$$

$$= 11.09.$$

$r=0,\ k=5$

b) $c = \chi^2_{4,0.95} = 9.488 < 11.02 = \chi^2$.

we reject the hypothesis.

c) $c = \chi^2_{4,0.975} = 11.14 > 11.02 = \chi^2$

we accept