ERASMUS B4 Course

Winter Semester 2020/2021, 14 January

Subject: Introduction to Mathematical Statistics (MS)

I. The concept of general population

Let's start by explaining the fundamental difference between probability theory and mathematical statistics. In both theories the concept of a random variable appears. So where is this difference?

In P.T. we assume that we know the distribution of X and on this basis we can calculate EX, $Var(X)$, probabilities $P(A)$, $A \in \Sigma$.

In M.S. the distribution of X is unknown and the main goal of M.S. is to determine this distribution.

Let's try to explain the reason for this situation. We will do it on an example.

Ex 1 - the PREDICTION PROBLEM

Let us take the elections phenomenon on the example of a country with a multimillion population of people. In such a situation, we can assume that the number of voters is also multi-million.

For this reason we can say that we are dealing with a mass (global) phenomenon.

The question is whether it is possible to predict with a high degree of certainty the result of the elections. If so, how can it be done?

We know the answer - by using methods of MS!

But let's go further, and ask why it is possible to solve such problem?

The answer is as follows - this possibility gives us a stochastic description of the behavior of the population - voters

Finally, let us ask where this description comes from?

The answer is as follows - from the adopted methodology of population description.

To be precisely, if by $\mathbb{P}$ we denote the population, then $p \in \mathbb{P} \iff p$ has some (the same) feature (characteristic), which is denoted by X.

Now, if we understand this feature well, we will able to predict the interactions in the whole population, and consequently, in our case to predict the vote results.

Of course, in M.S. X means a random variable and its probability distribution.

Def 1 (general population)

Let $\mathbb{P}$ denote the set. We say that $\mathbb{P}$ is a general population, if the following relation holds:

$$p \in \mathbb{P} \iff p \text{ has a feature } X.$$

Remarks.

In applications we have more than one feature. In this course we have decided to reduce problem only for a single feature.

TASK 1. For given examples of populations try to give examples of features, where:

$\mathbb{P}$ = „production process", „capital market",
„healthcare politics", „climatology".

Assumption.

Whenever we say that „we have a general population $\mathbb{P}$ with the feature X" it means that $\mathbb{P}$ is a large set (with respect of its power).

II Stochastical model of $\mathbb{P}$ & X

Because of mass character of $\mathbb{P}$, we apply the probability description of X. Simply, X is a random variable with some (unknown) probability distribution. Therefore the theoretical description of $\mathbb{P}$ is K.P.M. (Ω, Σ, P), and then

$$\Omega \ni \omega \longrightarrow X(\omega) \in \mathbb{R},$$

$$\forall t \in \mathbb{R} \quad \{\omega \in \Omega : X(\omega) < t\} \in \Sigma$$

$$P(\{\omega \in \Omega : X(\omega) < t\}) = F_X(t),$$

as we know from P.T.

To solve our problem, we have only one source of information - namely the whole population. So we have to observe the Π. But (by assumption and in application) Π is a large set — too large to realize such observation.

Therefore we have to choose the representation of Π, so to take

$$\Pi_0 = \{P_1, P_2 .., P_n\} \subset \Pi,$$

where: a) n is not too large

b) Π_0 "well represented" Π

For the model we have

$$\Omega_0 = \{\omega_1, \omega_2 .., \omega_n\} \subset \Omega, \text{ and}$$

as a result of observation of Π, by using its representation we obtain

the sequence of the real numbers

$$(*)\quad (X_1, X_2, \ldots, X_n) = (X(\omega_1), X(\omega_2), \ldots, X(\omega_n)),$$

so $X(\omega_j) = x_j$, $j = 1, 2, \ldots n$.

It can be proved, that the sequence $(*)$ which satisfies conditions (a) & (b) it can be constructed in a diffent way, namely

$$(**)\quad (X_1, X_2, \ldots X_n) = \underbrace{(X_1, X_2, \ldots, X_n)}_{\text{random vector (!)}}(\omega_0),$$

where

(i) $d(X_j) = d(X)$, $j = 1, \ldots, n$

(ii) $\{X_1, X_2, \ldots, X_n\}$ are S.I.

for some $\omega_0 \in \Omega$.

Def 2 (Simple Sample)

By the SIMPLE SAMPLE ~~of~~ corresponding to given general population $\mathcal{P}$ with the feature X

We understand the result of observation of $\vec{\Pi}$ which has the form

$$(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$$

with property (**), when (i) and (ii) hold

III 3 ways of presenting of the random observation of $\vec{\Pi}$.

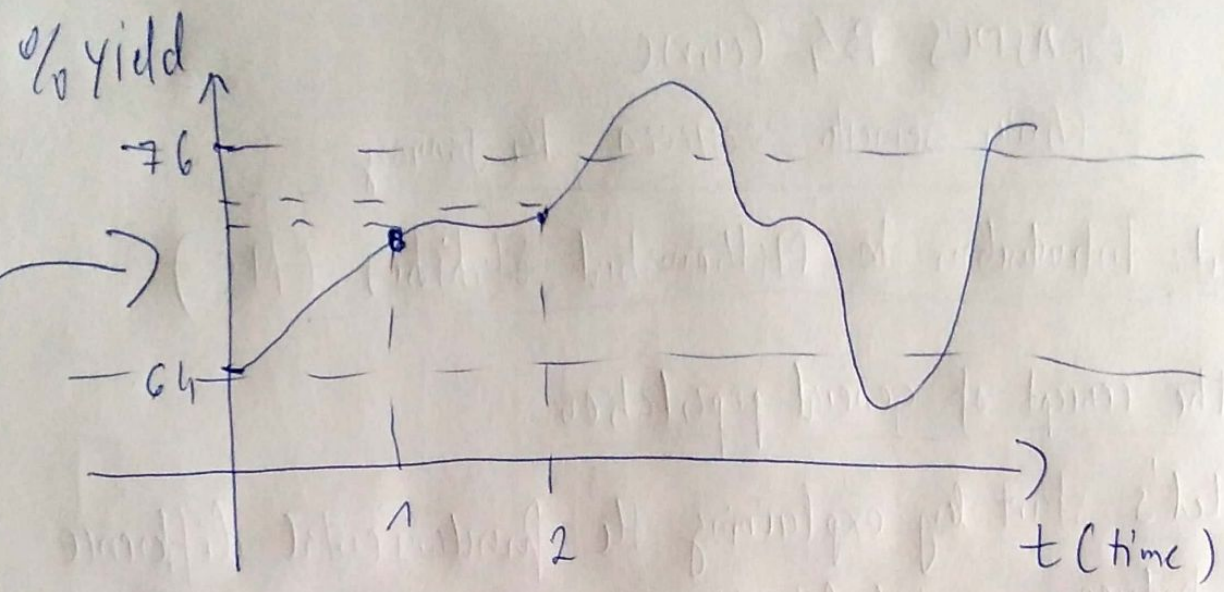
We will show now, how we can present the S.P. $(X_1, X_2, \dots, X_n) = (X_1, X_2, \dots, X_n)(\omega_0)$ corresponding to given G.P. $\vec{\Pi}$ with the feature X.

For this purpose we consider the following next example.

Ex 2.

Let us consider, for example, a chemical process that is producing batches of a desired material, where the feature X is the percentage yield in this process.

This process is observed as a function of time argument. So we can image, that the situation is as in the above graph



Suppose we observe the process 200 times in such a way, and the sample values change from 64% to 76%.

∇ (The results of all observations is included in attached table ~~on the end of this materials~~).

Dividing the range of observation (64%, 76%) into 12 equal intervals, and plotting the total number of observed yields in equal intervals as the height of rectangle over the interval results we can obtain in the can the following graph:

Table 8.1 Chemical yield data (data source: Hill, 1975)

Batch no.	Yield (%)	Batch no.	Yield (%)	Batch no.	Yield (%)	Batch no.	Yield (%)	Batch no.	Yield (%)
1	68.4	41	68.7	81	68.5	121	73.3	161	70.5
2	69.1	42	69.1	82	71.4	122	75.8	162	68.8
3	71.0	43	69.3	83	68.9	123	70.4	163	72.9
4	69.3	44	69.4	84	67.6	124	69.0	164	69.0
5	72.9	45	71.1	85	72.2	125	72.2	165	68.1
6	72.5	46	69.4	86	69.0	126	69.8	166	67.7
7	71.1	47	75.6	87	69.4	127	68.3	167	67.1
8	68.6	48	70.1	88	73.0	128	68.4	168	68.1
9	70.6	49	69.0	89	71.9	129	70.0	169	71.7
10	70.9	50	71.8	90	70.7	130	70.9	170	69.0
11	68.7	51	70.1	91	67.0	131	72.6	171	72.0
12	69.5	52	64.7	92	71.1	132	70.1	172	71.5
13	72.6	53	68.2	93	71.8	133	68.9	173	74.9
14	70.5	54	71.3	94	67.3	134	64.6	174	78.7
15	68.5	55	71.6	95	71.9	135	72.5	175	69.0
16	71.0	56	70.1	96	70.3	136	73.5	176	70.8
17	74.4	57	71.8	97	70.0	137	68.6	177	70.0
18	68.8	58	72.5	98	70.3	138	68.6	178	70.3
19	72.4	59	71.1	99	72.9	139	64.7	179	67.5
20	69.2	60	67.1	100	68.5	140	65.9	180	71.7
21	69.5	61	70.6	101	69.8	141	69.3	181	74.0
22	69.8	62	68.0	102	67.9	142	70.3	182	67.6
23	70.3	63	69.1	103	69.8	143	70.7	183	71.1
24	69.0	64	71.7	104	66.5	144	65.7	184	64.6
25	66.4	65	72.2	105	67.5	145	71.1	185	74.0
26	72.3	66	69.7	106	71.0	146	70.4	186	67.9
27	74.4	67	68.3	107	72.8	147	69.2	187	68.5
28	69.2	68	68.7	108	68.1	148	73.7	188	73.4
29	71.0	69	73.1	109	73.6	149	68.5	189	70.4
30	66.5	70	69.0	110	68.0	150	68.5	190	70.7
31	69.2	71	69.8	111	69.6	151	70.7	191	71.6
32	69.0	72	69.6	112	70.6	152	72.3	192	66.9
33	69.4	73	70.2	113	70.0	153	71.4	193	72.6
34	71.5	74	68.4	114	68.5	154	69.2	194	72.2
35	68.0	75	68.7	115	68.0	155	73.9	195	69.1
36	68.2	76	72.0	116	70.0	156	70.2	196	71.3
37	71.1	77	71.9	117	69.2	157	69.6	197	67.9
38	72.0	78	74.1	118	70.3	158	71.6	198	66.1
39	68.3	79	69.3	119	67.2	159	69.7	199	70.8
40	70.6	80	69.0	120	70.7	160	71.2	200	69.5

given in Table 8.2 are six-year accident records of 7842 California drivers (Burg, 1967, 1968). Based upon this set of observations, the histogram has the form given in Figure 8.2. The frequency diagram is obtained in this case simply by dividing the ordinate of the histogram by the total number of observations, which is 7842.

number of accidents per driver during a six-year time span in Canada.

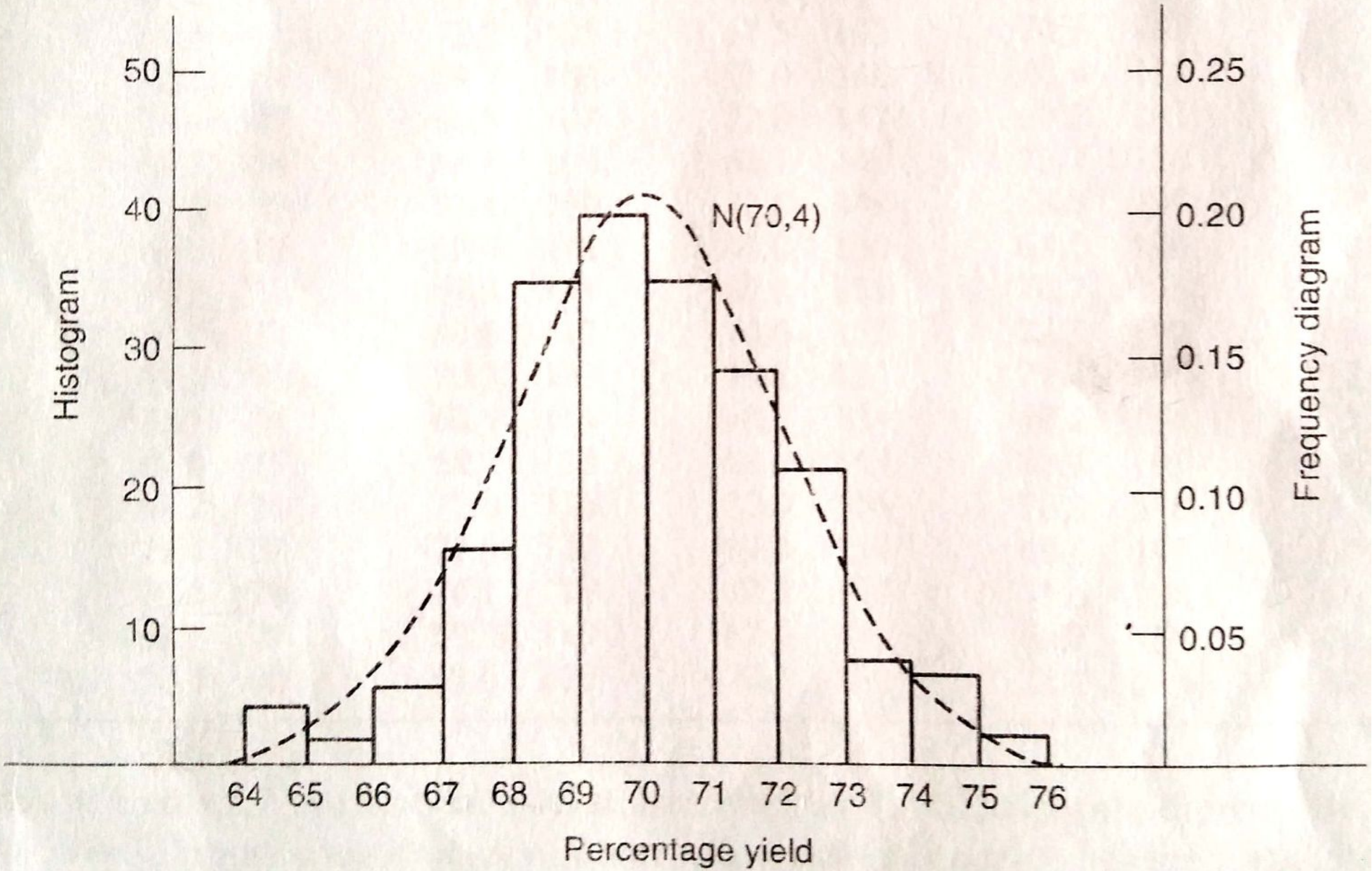


Figure 8.1 Histogram and frequency diagram for percentage yield

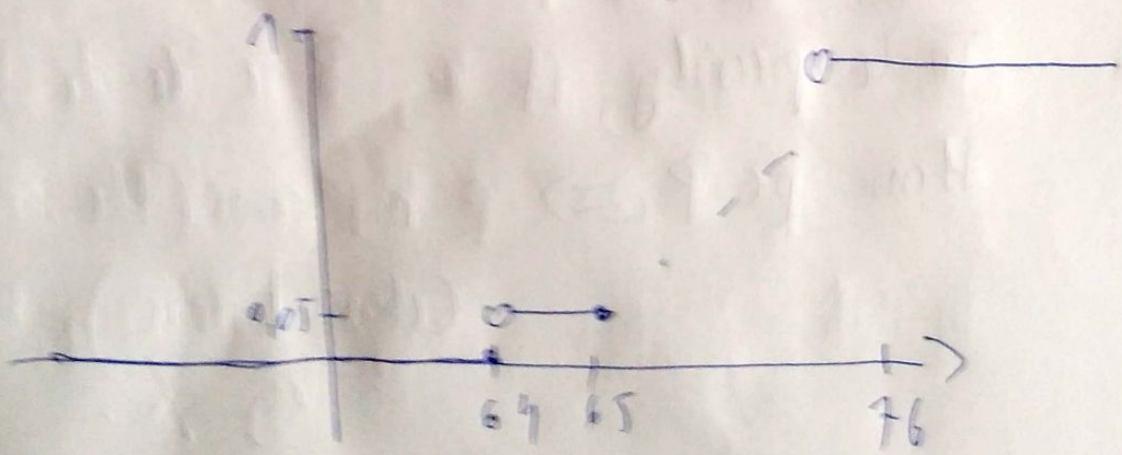
the type of presentation of S.S. is called HISTOGRAM

At the same time, we can compute so called FREQENCY DIAGRAM as a result of transformation $\frac{\#\text{ of observed in a given interval}}{\#\text{ observations } (= 200)}$

(look at the scale on the right side of the grid of HISTOGRAM).

Finally, the frequency diagram allows us to plot the third type of presentation of S.S. which is called the EMPIRICAL C.D.F., remarking the function

$$\mathbb{R} \ni t \longrightarrow \frac{\#\{x_i : x_i \leq t\}}{\#\text{ of observations}}$$



Results of observation based on histogram and others

1) Histogram suggests, that

$X \in N(m, \sigma^2)$, because we see the bell curve.

2) so

$$f_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-m)^2}{2\sigma^2}}, \quad t \in \mathbb{R}$$

From the frequency diagram follows that

$$f_X(70) = 0{,}2, \quad m \cong 70[\%]$$

$$\Downarrow$$

$$\frac{1}{\sqrt{2\pi}\sigma} = 0{,}2 \Rightarrow \sigma = \frac{5}{\sqrt{2\pi}} \cong 2$$

Finally, $X \in N(70, 2^2)$

Remarks:

1°. The two plots (Histogram and E.C.D.F.) are similar to the two functions used in P.T.:

histogram — density function (f_X)
E.C.D.F. — C.D.F. (F_X).

2°. Formally, at this moment, we do not have to right to claim, that the presented results in the form of histogram, frequency diagram or E.C.D.F. describe the feature of X in terms of its distribution, because they are related only to 200 observations

3°. These results (at this moment) only are related to S.S.

4°. M.S. can extrapolate the results related to S.S. on the whole general

population Π.

On the next lecture I am going to show you this extrapolation procedure.