ERASMUS BY Course

Winter Semester 2020/2021, 21 JANUARY

Subject: topics in statistical inference for given general population with the feature X.

Introduction.

Let the general population Π with the feature X be given.

We know that the main goal of M.S. is to establish the probability distribution of X.

To do this, we need the record of observation of Π. It means, that we have to choose from Π its representation Π_0 and examine it.

In M.S. we assume that the result of the above procedure has a form

$$\mathbb{R}^n \to (X_1, X_2, \dots, X_n) = (X_1, X_2, \dots, X_n)(\omega_0),$$

where $d(X_j) = d(X)$, $j = 1, \dots, n$,

and $X_1, \dots, X_n$ are stoch. indep.

then we say that we have the SIMPLE SAMPLE.

The last lecture showed us that the presentation of S.S. is very important in statistical inference.

It is clear that S.S. describes $\mathbb{P}_0 \subset \mathbb{P}$. We are going to explain below that the information for $\mathbb{P}_0$ can be extrapolated to the whole $\mathbb{P}$.

The concept of Empirical Cummulative Prob. Dist (ECPD)

Fix the S.S.

$$(X_1, X_2, \dots, X_n) = (X_1, X_2, \dots, X_n)(\omega_0).$$

The function

$$\mathbb{R} \ni t \longrightarrow F_n(t, \omega_0) \overset{\text{def}}{=} \frac{1}{n} \#\{i : X_i < t\}$$

is called ECPD corresponding to $\mathbb{P}$ with X for given S.S.

We know, that ECPD we can obtain from histogram (or frequency diagram) and vice versa. Moreover, the graph of ECPD is a "step function".

Why ECPD is so important in M.S. ?

It is explained by the following theorem

Theorem (on ECPD) (GNIEDENKO)

For given G.P. Π with the feature X let us consider the sequence of functions

$$\mathbb{R}\times\Omega \ni (t,\omega) \longrightarrow F_n(t,\omega) \stackrel{df}{=} \frac{1}{n}\#\{i : X_i(\omega) < t\},$$

where for $n \geq 2$

$X_1, X_2, \dots, X_n$ are st. independent with the same distribution X.

If F_X is (unknown) a cummulative prob. dist. f. corresponding to X, then for the events

$$A_t = \{\omega\in\Omega : F_n(t,\omega) \longrightarrow F_X(t)\}, \quad t\in\mathbb{R}$$

we have

$$P(A_t) = 1, \quad t\in\mathbb{R}.$$

Remarks

1°. Each member $F_n(t,\omega)$ of the above sequence is a generalization of ECDF, namely

$$F_n(t,\omega_0) = F_n(t,\omega)\Big|_{\omega=\omega_0}.$$

So if we take $(F_n(t,\omega))_{n\geq 1}$ it means we consider (from theoretical point of view!) all S.S. (with respect of n and $\omega\in\Omega$).

2°. Since, for given t, $P(A_t)=1$, we can assume that ω_0 from the definition of S.S. belongs to A_t.

Therefore we have an approximation rule

$$\boxed{F_n(t,\omega_0) \simeq F_X(t)}$$

which is the basic argument in extrapolation procedure.

The concept of the estimation

The stage of the statistical inference using the concept of ECDF generally does not solve the main problem of M.S.

On the output of this stage the typical solution is as follows:

$F_X(t, \theta)$ - we know the „shape" of F_X, but the value of the parameter θ is unknown.

So, we need the next stage of S.I, to ~~estimate~~ establish the value θ.

It is called the ESTIMATION PROCEDURE.

The main idea of E.P. says:

For given $\alpha \in (0,1)$ (significant level), for instance $\alpha = 0{,}05$ or $0{,}1$

We try to construct of two r.v. Z_1, Z_2 such that:

(i) Z_1, Z_2 is constructed by using the random vector $(X_1 \dots X_n)$ which defines the S.S.

(ii) $Z_1(\omega) < Z_2(\omega)$ with probability 1

(iii) $P(\text{Event: } Z_1(\omega) < \theta < Z_2(\omega)) = 1 - \alpha$.

Then we say we have a random interval (Z_1, Z_2).

Now, if we use $\omega \in \Omega$ from the S.S.

We obtain the real interval $(Z_1(\omega_1), Z_2(\omega_1)) \subset \mathbb{R}$,

and finally (by (iii)) we have the following conclusion

$$\theta \in (Z_1(\omega_1), Z_2(\omega_1)) \text{ with prob. } 1 - \alpha$$

Example

Suppose that we know (for instance from analysis of histograms) that the S.I.

$$(x_1, \dots, x_n) = (X_1, \dots, X_n)(\omega), \text{ that}$$

$X \sim N(m, \sigma^2)$, and the value σ^2 is fixed.

If we fix $\alpha \in (0,1)$ and we take

$$Z_1 \stackrel{\text{def}}{=} \bar{X}_n - u_\alpha \frac{\sigma}{\sqrt{n}}, \quad Z_2 = \bar{X}_n + u_\alpha \frac{\sigma}{\sqrt{n}},$$

where $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$, we can find u_α if we resolve the equation

$$(\#) \quad P(\{\omega \in \Omega : Z_1(\omega) < m < Z_2(\omega)\}) = 1 - \alpha.$$

Namely,

$$(\#) \iff P(\{\omega \in \Omega : \bar{X}_n(\omega) - u_\alpha \frac{\sigma}{\sqrt{n}} < m < \bar{X}_n(\omega) + u_\alpha \frac{\sigma}{\sqrt{n}}\}) = 1 - \alpha$$

$$\iff$$

Proven: $P\left(-z_\alpha < \frac{\bar{X}_n(\omega) - m}{\frac{\sigma}{\sqrt{n}}} < z_\alpha\right) = 1-\alpha$.

But, according to our assumption of X,

$$\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \in N(0,1), \text{ so}$$

$$(\#) \iff \Phi(z_\alpha) - \Phi(-z_\alpha) = 1-\alpha \iff$$

$$2\Phi(z_\alpha) - 1 = 1-\alpha \iff$$

$$\boxed{\Phi(z_\alpha) = 1 - \frac{\alpha}{2}}$$

For instance, for $\alpha = 0{,}05$, $z_\alpha = 1{,}96$

The above shows, that by using z_1, z_2 given as above, we can construct the random interval and finally to realize E.S.

Let's realize the following numerical simulation.

→ $(x_1, x_2, x_3, x_4) = (2{,}20, 2{,}32, 1{,}97, 2{,}01)$ /

for $X \in N(m, (0{,}25)^2)$ and $\alpha = 0{,}05$.

Since for $\alpha = 0{,}05$, $u_\alpha = 1{,}96$, we have

$$Z_1(v) = \overline{X}_4(v) - 1{,}96 \frac{0{,}25}{2}$$

$$Z_2(v) = \overline{X}_4(v) + 1{,}96 \frac{0{,}25}{2} \quad , v \in V.$$

So, for $v = v_0$

$\overline{X}_4(v_0) = 2{,}08$ and

$$Z_1(v_0) = 2{,}08 - 1{,}96 \frac{0{,}25}{2} = 1{,}838$$

$$Z_2(v_0) = 2{,}08 + 1{,}96 \frac{0{,}25}{2} = 2{,}328.$$

Thus, in this case

$m \in (1{,}838, 2{,}328)$ with prob. $0{,}95$

Conclusion :

The ideas presented above belong to many others that make up the stochastical inference procedure (S.I.P).

The main goal of S.I.P. is to establish the probability distribution of the feature X for given general population P.

But to present them methods, you need another course that I invite you to take in the future.

PR