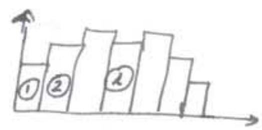
STATISTICS IN HYDROLOGY .....

Yesterday, we discussed on



→ Relative frequency function $f_s(x_i)$

→ The probability estimate in the interval $[x_i - \Delta x, x_i]$

$$P(x_i - \Delta x \leq x \leq x_i)$$

→ The cumulative frequency function $F_s(x_i)$

$$F_s(x_i) = \sum_{j=1}^{i} f_s(x_j)$$

→ Equivalent $P(X \leq x_i)$ the cumulative probability

$$\lim_{\substack{n \to \infty \\ \Delta x \to 0}} \frac{f_s(x_i)}{\Delta x} = f(x) = \text{probability density function}$$

$$F(x) = \text{Probability distribution function}$$

$$= \lim_{\substack{n \to \infty \\ \Delta x \to 0}} F_s(x)$$

Also $f(x) = \dfrac{d}{dx} F(x)$

→ Again the meaning of $F(x)$ and $F_s(x_i)$ can be seen through Figure 11.2.1

→ We also described about a probability density function

→ the normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

where μ and σ are parameters.

The standard normal distribution (again a probability density function)

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad ; \quad -\infty \leq z \leq \infty$$

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \, du$$

where $u \rightarrow$ dummy variable.

The Statistical Parameters

Q: We asked you on what is the objective of statistics.

→ To extract essential information from a set of data

→ To reduce large set of numbers to a small set.

Parameters like → mean, standard deviation, skewness, etc.

Q: What is a statistical parameter?

If X is the concerned random variable, then a statistical parameter is the expected value E of ~~the~~ some function of the random variable.

→ The most commonly used expected value is the expected value of the random variable itself.

This is mean.

i.e. $$E[x] = \mu = \int_{-\infty}^{\infty} x f(x)\, dx$$

$E[x]$ is the first moment about the origin of the random variable. It describes central tendency.

If you are having a sample $x_1, x_2, x_3, \ldots x_n$ for X the Sample Average $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$

Refer Table 11.3.1 in the text book.

→ Variability σ^2 = second moment about mean

$$E\left[(x-\mu)^2\right] = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

→ Symmetery of a distribution (i.e probability density function) is measured using skewness.

$$E\left[(x-\mu)^3\right] = \int_{-\infty}^{\infty} (x-\mu)^3 f(x)\, dx$$

(Third moment about mean).

Coefft. of skewness, $\gamma = \dfrac{1}{\sigma^3} E\left[(x-\mu)^3\right]$

To fit Probability Distribution

You know by now that probability distribution is a function describing the probability of occurrence of the random variable.

→ There will be a set of hydrologic data available

→ If you now fit an appropriate probability distribution to the given data → the sample can be now represented through its parameters - like mean, variance, skewness, etc.

Method of Moments

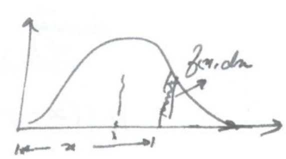
The probability distribution will be more accurate if the first moment calculated using sample data

i.e. $\bar{x} = \frac{1}{n}\left(x_1 + x_2 + x_3 + \cdots + x_n\right)$

and the first moment of the assumed probability density function $f(x)$ about the origin are equal.

i.e. $E(x) = \int_{-\infty}^{\infty} x\, f(x)\, dx$

Actually $\int_{-\infty}^{\infty} x\, f(x)\, dx$ = Centroid of the area described by probability density function.



$$\mu = \int_{\infty}^{\infty} x\, f(x)\, dx.$$

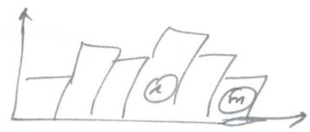
→ Similarly the second moments and third moments are also compared.

Goodness of Fit

We can also appropriately fit probability distribution to a given sample data by goodness of fit approach.

→ We can test the relative frequency functions and cumulative frequency functions of the sample as well as theory.

→ In sample $f_s(x_i) = \frac{n_i}{n}$



The theoretical value for this is $P(x_i - \Delta x \leq X \leq x_i) = p(x_i)$

$$= F(x_i) - F(x_{i-1})$$

Using chi χ^2-test we can perform the goodness of fit.

Define (a variable) $$\chi_c^2 = \sum_{i=1}^{m} \frac{n\left[f_s(x_i) - p(x_i)\right]^2}{p(x_i)}$$

$m \rightarrow$ number of intervals.

Again $n f_s(x_i) = n_i$

$n\,p(x_i) \rightarrow$ expected number of occurrences in interval i.

We have to now define χ^2 probability distribution.

χ^2 distribution with ν degrees of freedom is the distribution for the sum of squares of ν independent standard normal variable z_i.

$$\chi_\nu^2 = \sum_{i=1}^{\nu} z_i^2$$

In the χ^2 test, we have $\nu = m - p - 1$

$m \rightarrow$ no: of intervals
$p \rightarrow$ no: of parameters used in the proposed distribution

* You chose a confidence level $1 - \alpha$.
$\alpha \rightarrow$ significance level

Null Hypothesis $\rightarrow$ Proposed probability distribution fits the data.

$\rightarrow$ Hypothesis rejected if $\chi_c^2 >$ limiting value.

Some of the probability distributions

Normal distribution
Lognormal
Exponential
Gamma.

Normal distribution $\rightarrow$ based on central limit theorem

If there X_i independent random variables, but they have mean μ and variance σ^2, then the distribution of sum of l such random variables

i.e. $Y = X_1 + X_2 + \cdots + X_l$

Then mean will be $\rightarrow l\mu$
Variance $\rightarrow l\sigma^2$
$\}\rightarrow$ Normal distribution

This is applicable irrespective of the type of probability distribution of each individual RV.

In sample data of n - observations

$$\bar{x} = \frac{1}{n}\left[x_1 + x_2 + \cdots + x_n\right]$$

This sample data mean can then be approximated as normal mean μ

Hydrologic variables → annual precipitation follow normal dist.

→ Most hydrologic data maybe non-negative.

Exponential distribution used in streamflow data analyses.

Gamma distribution used in analysing depth of storms, etc.

FREQUENCY ANALYSIS

We have extreme events like

→ Huge thunderstorms
→ Cyclones
→ Droughts
→ Floods

The magnitude of such events can vary considerably.

Q: Now how frequent can such events occur?

Usually very severe of those events occur less frequently.

So now we can relate the magnitude of extreme events and their frequency of occurring. Frequency Analysis.

→ For design of bridges, culverts, dams etc. we require the interpretations from frequency analysis.

Return Period

Consider maximum discharge in a year from a river. The observations are recorded say for 44 years as shown in Table 12.1.1 of the text book.

→ In the region, it is suggested that discharge above 50,000 cfs is severe.

→ Then we want to find return period of flood $\geq$ 50,000 cfs.

→ From Fig 12.1.1, we can see 9 years where flood or discharge exceeded 50,000 years.

Return period T of the event $X \geq x_T$

$(x_T = 50{,}000$ cfs here)

is the expected value of τ itself

i.e. $E(\tau)$,

where τ → recurrence interval between occurrence of the event $X \geq x_T$.