Basic Statistics: Concepts of Distribution, Correlation, regression and hypothesis testing

Somenath Dutta
Office of Climate Research & Services

भारत मौसम विज्ञान विभाग
INDIA METEOROLOGICAL DEPARTMENT

Concept of distribution

- **Types of variables:**
  - **Continuous: Can have any real value, for e rainfall, temperature, pressure, humidity, w**
  - **Discrete: Can have only discrete values, f Number of rainy days, number of cyclones heat wave days etc.**

Suppose X is a random variable. If it is a dis variable, then it can have some discrete valu

Value or range of values of the variable 'X'	Probability that 'X' assumes a real or a range of values 'X' can assume
$(X = x_r)$ or $(-\infty < X \ll x)$	$P(X = x_r)$ *or* $P(-\infty < X \ll$

$X = x$

if X is

value up to say 'x'. Then $(-\infty < X \ll x)$ can

Example:

Number of rainy days in 2017 SWMS	Frequency	Probability
1	f_1	$p_1 = f_1 / \sum_1^n f_r$
2	f_2	$p_2 = f_2 / \sum_1^n f_r$
.	.	.
.	.	.
r	f_r	$p_r = f_r / \sum_1^n f_r$
.	.	.
.	.	.
n (large)	f_n	$p_n = f_n / \sum_1^n f_r$
total	$=\sum_1^n f_r$	

Distribution functions:

- For discrete variables:
  - Probability mass function
  - cumulative distribution function.
- For continuous variables:
- Probability density function
- probability distribution function

Distribution functions

- **PMF & CDF: For a discrete variable, X, PMF i**

  $p_X(x_r) = P(X = x_r)$ **and the CDF is given as** P

  $\sum_{r=-\infty}^{r=k} p_x(x_r) = \sum_{r=-\infty}^{r=k} P(X = x_r)$.

- **PDF & CDF: For a continuous variable 'X', th given by**

$F_X(x) = P(-\infty < X \ll x)$ **and PDF is given by** f_X

$$\lim_{h\to 0} \frac{F_X(x+h)-F_X(x)}{h} = \lim_{h\to 0} \frac{P(-\infty<X\ll x+h)-P(-\infty<X\ll x)}{h} =$$

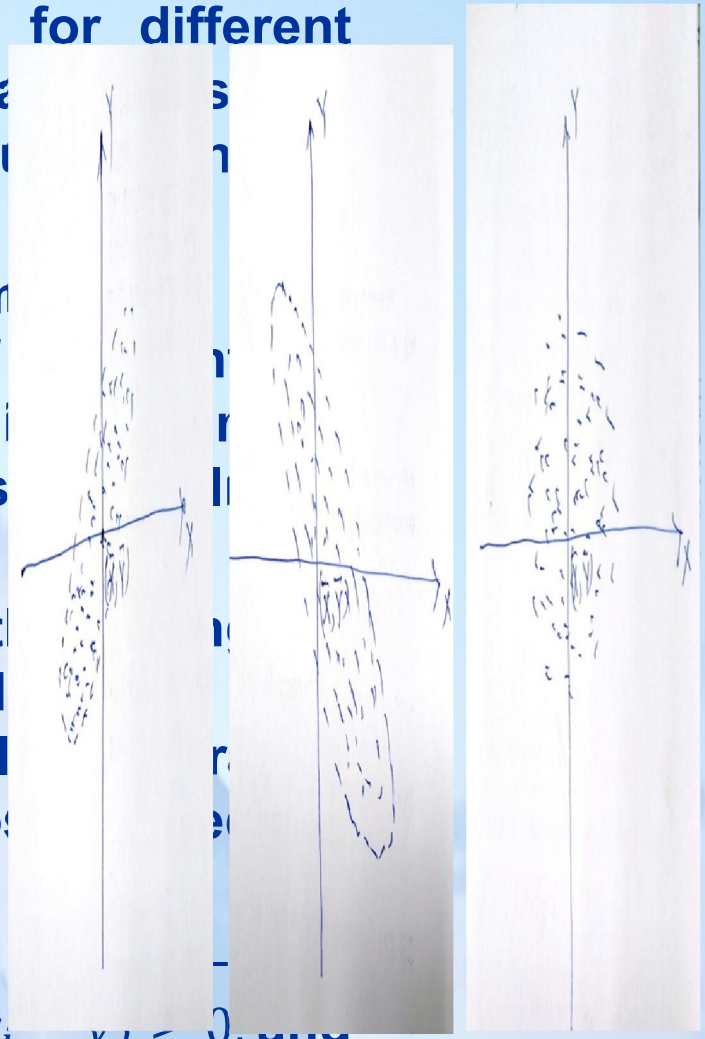
- **Some properties of CDF:**
  - $P_X(-\infty) = 0$
  - $P_X(\infty) = 1$
  - $F_X(-\infty) = 0$

Correlation analysis

- In Meteorology, one of the important forecast technique is Statistical technique. In this technique an unknown variable, say, ‘Y’, is attempted to forecast using a known variable, say, ‘X’, by establishing a suitable functional relation between them.
- However, for that one has to first ensure that the y is associated with X, which can be established from previous values of them.
- There are many simple methods to examine the association between two variables, viz., scatter diagram, co-variance and correlation.

•In the adjoining figures, along x-axis independent variable (x) and along y-axis dependent variable (y) are measured. Values of dependent variable for different values of independent variable, are plotted a s (x,y), in the x-y plane. Assemble of all su n called scatter diagram.

•In the adjoining figures 3 scatter diagram shown. In the 1st one it is seen that most of n in I and III quadrants, in the 2nd one they li r quadrants where as in the 3rd one points l uniformly distributed in all quadrants.



•For points in I and III quadrants, indicates t n both variables are in the same d $(x-\bar{x})(y-\bar{y}) \gg 0$. For points in II and I r indicates that change in x, y are in the oppo e and $(x-\bar{x})(y-\bar{y}) \leq 0$.

•Thus, for the 1st scatter diagram, $\sum_1^n (x_i$ - 0, for the 2nd scatter diagram, $\sum_1^n (x_i - \bar{x})(y_i$ $y) \simeq 0$, and for 3rd scatter diagram, $\sum_1^n (x_i - \bar{x})(y_i - \bar{y}) \approx 0$.

•Hence, $\sum_1^n (x_i - \bar{x})(y_i - \bar{y})/n$, can be taken as a measure of association between x and y. This is known as covariance

- Limitation with covariance is that, it is not n dimensional. So, difficult to compare 2 or m variances.
- To get rid of this limitation, another quantity normalizing the covariance by the product o Y. This quantity is called correlation coefficie between x and y. Thus , $CC = \frac{cov(x,y)}{\sigma_x \sigma_y}$.
- Properties of CC:
  - CC is dimension less, it doesn't have any purely a number.
  - $-1 \leq cc \leq 1$
  - If CC= $+1$, then we say, X,Y are perfectly co

Simple linear regression analysis

- Let us consider a bi-variate variable (X,Y), wh dependent variable Y, changes with change i independent variable, X.
- Suppose, we have a time series of X and Y fo period. Generally, it is desired to predict Y fo period for any future value of X.
- In simple linear regression analysis, it is pres linear relation like, Y=aX+b, exists between X a,b are parameters determined using princip square.
- It can be shown that the equation of the requ

Test of hypothesis

- We know that while carrying an experiment we have three stages, viz., Experiment, Observation, Inference.
- Similarly while carrying a statistical experiment, first we conduct collection of samples, then analysis of sample data and finally inference.
- Statistical inference has two parts, viz., estimation and hypothesis testing.
- Estimation is beyond the scope of our discussion.
- Test of hypothesis is carried out to establish an existing conjecture or to reject the same, based on some random sample.

Examples of hypothesis testing

- For example, say, based on 1951-1980 data, we know that number of Western disturbances passing Indian longitude is ‘n’. We want to examine whether in the recent period (1981-2010), number of WDs passing Indian longitude has remain ‘n’ or has changed.
- Similarly, say, based on some old data set we know that, number of LOPAR in SWMS intensifying to MD is ‘n’. We want to verify whether this number has changed significantly or not in the recent time.
- Another example, say, based on old sample data, it is known that MOK is 1st June. We want to verify whether date of MOK, in recent time, has changed significantly or not.

Some useful terms in Hypothesis testing

- Null hypothesis: An hypothesis to be tested whether it or rejected. We denote it as H_0.
- Alternate hypothesis: an hypothesis against which the tested. It is denoted by H_1 .
- Test of a Hypothesis : A rule telling us when a null hyp or reject.
- Test statistic: This is the statistic used in formulating a
- Error in test: Based on a random sample drawn fr ultimately one concludes, H_0 is accepted or rejected. S happen that a true hypothesis is rejected or a fal accepted. Obviously these are errors in the test. They called type I error and type II error.
- Critical region: It is a subset of the sample space, rejected, i.e., C=$\{x \in S / H_0$ is rejected$\}$, where S is the sa is critical region.
- Level of significance: This is the probability level, u employed in defining the critical region. It is generally is customarily taken to be 0.05 or 0.01 etc.

How a test is selected?

- To conduct any statistical test about parameters of a population,
  - first a random sample is drawn from the population
  - Then null hypothesis is set up against an alternative hypothesis.
  - Then level of significance, i.e., the probability of committing type I error is set up.
  - For a given level of significance, there exists many number of tests, as such theoretically there exists an infinite number of tests for a given null hypothesis and level of significance.
  - We need to find out that test for which probability of committing type II error is least, i.e., power is maximum.
  - That test is the best test for which power is maximum.

Different types of test and corresponding test statistic

Name of the test	Test statist
Test of a mean (μ) of a population, when the SD (σ) of the population is known.	$H_0: \mu = \mu_0 \; against \; H_1: \mu > \mu_0 \; or$ $z = \frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma}$. For 1st case, H_0 is rejected rejected If $Z < -Z_\epsilon$ and finally for 3rd cas
Test of a mean (μ) of a population, when the SD of the population is not known.	$H_0: \mu = \mu_0 \; against \; H_1: \mu > \mu_0 \; or$ $t = \sqrt{(n-1)}\frac{\bar{x}-\mu_0}{s}$. For 1st case, H_0 is rej H_0 is rejected If $t < -t_{\epsilon,n-1}$ and finally f $\lvert t\rvert > t_{\epsilon/2,n-1}$.
Comparison of two means of two normal populations, based on random samples of sizes n_1 and n_2 *drawn from them, with*known standard deviations.	$H_0: \mu_1 = \mu_2 \; against \; H_1: \mu_1 > \mu_2 \; or$ $Z = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$. For 1st case, H_0 is reje is rejected If $Z < -Z_\epsilon$ and finally for 3rd
Comparison of two means from two normal populations, based on random samples of sizes n_1 and n_2 *drawn from them, with un*known population standard deviations.	$H_0: \mu_1 = \mu_2 \; against \; H_1: \mu_1 > \mu_2 \; or$ $Z = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$. For 1st case, H_0 is reje case H_0 is rejected If $t < -t_{\epsilon,n_1+n_2-2}$ and rejected If $\lvert t\rvert > t_{\epsilon/2,n_1+n_2-2}$.

Different types of test and corresponding test statistic

Name of the test	Test statist
Comparison of k means of k normal populations, based on random samples of sizes $n_1, n_2, \ldots\ldots, n_k$, *drawn from them, with*common known commonstandard deviationσ,	$H_0: \mu_1 = \mu_2 = \cdot\cdot$ *against all alternatives on the ba* *Of sizes* $n_1, n_2, \ldots\ldots, n_k$. $MSB = \frac{\sum_1^k n_i(\bar{x}_i - \bar{x})^2}{(k-1)}$, $MS\ W = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}{(n-k)}$ rejected If $F > F_{\epsilon;k-1,n-k}$.
Comparison of two variances of two normal populations, based on random samples of size n_1 and n_2, *drawn from them*.	$H_0: \sigma_1 = \sigma_2\ against\ H_1: \sigma_1 > \sigma_2$ or $F = \frac{s_1^2 n_1(n_2-1)}{s_2^2 n_2(n_1-1)}$. For 1st case, H_0 is reject case H_0 is rejected If $F < F_{1-\epsilon;\, n_1-1,n_2-$ rejected If $F > t_{\epsilon/2;n_1-1,n_2-1}$.
Test of significance of correlation coefficient	$H_0: \rho = 0, against\ H_1: \rho > 0$ or **t=**$\frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$**. For 1st case,** H_0 **is rejected If** t **rejected If** $t < -t_{\epsilon,n-2}$ **and finally for 3r** $t_{\epsilon/2,n-2}$.

Thanks for patience hearing