www.ranjan.net.np

Probability, Continuous Random Variables, Statistics and descriptive statistics, Joint Variation of Two Variables

Ranjan Kumar Dahal, PhD, PostDoc, M.ASCE

Associate Professor,
Geodisaster Research Center, Central Department of Geology, Tribhuvan University, Kirtipur, Kathmandu, Nepal

2nd Lecture

www.ranjan.net.np

Course Contents – first part

- Statistics in Geology, Measurement Systems, Elementary Statistics
- Probability, Continuous Random Variables, Statistics and descriptive statistics, Joint Variation of Two Variables, Induced Correlations, Log ratio Transformation, Comparing Normal Populations, Central Limits Theorem , Testing the Mean P-Values, Significance, Confidence Limits, the t-Distribution, degrees of freedom, confidence intervals based on t, A test of the equality of two sample means, the t-test of correlation, The F-Distribution, F-test of equality of variances, Analysis of variance, Fixed, random, and mixed effects, Two-way analysis of variance, Nested design in analysis of variance, The Chi square Distribution, Goodness-of-fit test ,
- The Logarithmic and Other Transformations, Nonparametric Methods, Mann-Whitney test, Kruskai-Wallis test, Nonparametric correlation, Kolmogorov-Smirnov tests, Exercises.

Bayes' theorem

- *Bayes' theorem, named for Thomas Bayes, an eighteenth century English* clergyman who investigated the manner in which probabilities change as more information becomes available. Bayes' basic equation is:
- p(A,B) = p(B|A).p(A)



$$p(A,B) = p(B|A).p(A)$$

- which states that p(A,*B*), *the joint probability that both events A and B occur, is* equal to the probability that *B will occur given that A has already occurred, times* the probability that A will occur.
- p(BIA) is a conditional probability because it expresses the probability that *B will occur conditional upon the circumstance that* A has already occurred. If events A and *B are related (or dependent), the fact that* A has already transpired tells us something about the likelihood that *B will then* occur. Conversely, it is also true that

$$p(A,B) = p(A|B)p(B)$$



Finally

$$p(B|A)p(A) = p(A|B)p(B)$$

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

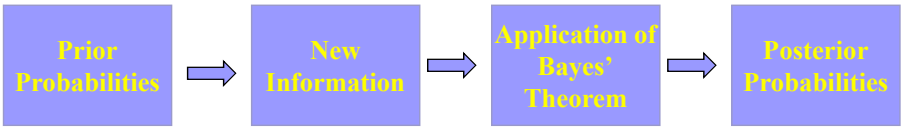
- If there is an all-inclusive number of events B_i *that are conditionally related to* event A, the probability that event A will occur is simply the sum of the conditional probabilities $p(A|B_i)$ times the probabilities that the events B_i *occur. That is,*

$$p(A) = \sum_{i=1}^{n} p(A|B_i)p(B_i)$$

substituted for p(A) in Bayes' theorem, the more general equation will be

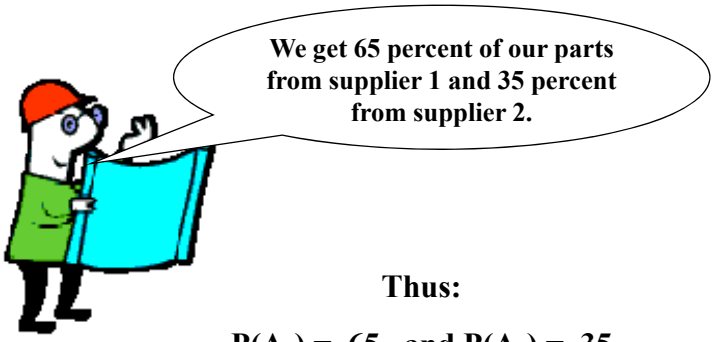
$$p(B_i|A) = \frac{p(A|B_i)\,p(B_i)}{\sum_{i=1}^{n} p(A|B_i)\,p(B_i)}$$

Probability Revision using Bayes' Theorem



Application of Bayes' Theorem

- Consider a manufacturing firm that receives shipment of parts from two suppliers.
- Let A_1 denote the event that a part is received from supplier 1; A_2 is the event the part is received from supplier 2



Thus:

$P(A_1) = .65$ and $P(A_2) = .35$

Quality levels differ between suppliers

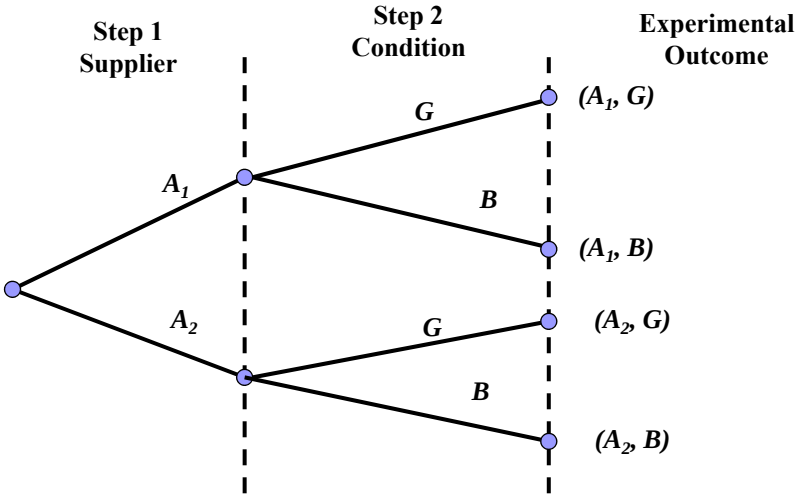
	Percentage Good Parts	Percentage Bad Parts
Supplier 1	98	2
Supplier 2	95	5

Let *G* denote that a part is good and *B* denote the event that a part is bad. Thus we have the following conditional probabilities:

$P(G \mid A_1) = .98$ and $P(B \mid A_2) = .02$

$P(G \mid A_2) = .95$ and $P(B \mid A_2) = .05$

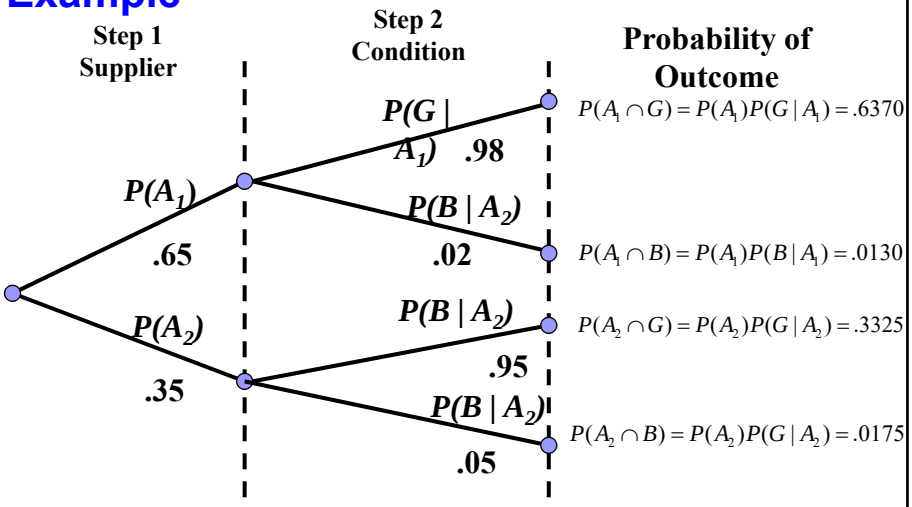
Tree Diagram for Two-Supplier Example

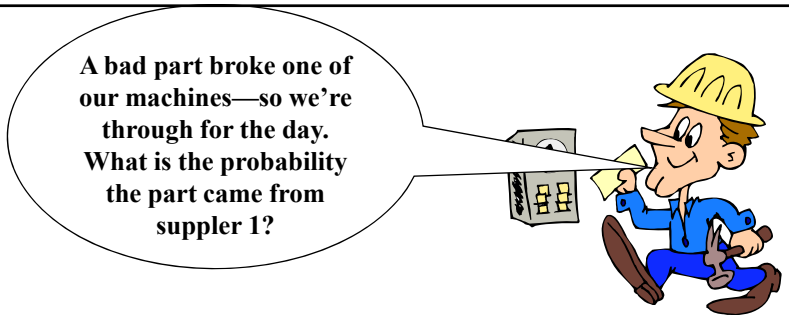


Each of the experimental outcomes is the intersection of 2 events. For example, the probability of selecting a part from supplier 1 that is good is given by:

$$P(A_1, G) = P(A_1 \cap G) = P(A_1)P(G \mid A_1)$$

Probability Tree for Two-Supplier Example





We know from the law of conditional probability that:

$$P(A_1 \mid B) = \frac{P(A_1 \cap B)}{P(B)} \qquad \textbf{(4.14)}$$

Observe from the probability tree that:

$$P(A_1 \cap B) = P(A_1)P(B \mid A_1) \qquad \textbf{(4.15)}$$

The probability of selecting a bad part is found by adding together the probability of selecting a bad part from supplier 1 and the probability of selecting bad part from supplier 2.

That is:

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B)$$
$$= P(A_1)P(B \mid A_1) + P(A_2)P(B / A_2) \qquad \textbf{(4.16)}$$

Bayes' Theorem for 2 events

By substituting equations (4.15) and (4.16) into (4.14), and writing a similar result for *P(B | A_2)*, we obtain Bayes' theorem for the 2 event case:

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)}$$

$$P(A_2 \mid B) = \frac{P(A_2)P(B \mid A_2)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)}$$

Do the Math

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)}$$

$$= \frac{(.65)(.02)}{(.65)(.02) + (.35)(.05)} = \frac{.0130}{.0305} = .4262$$

$$P(A_2 \mid B) = \frac{P(A_2)P(B \mid A_2)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)}$$

$$= \frac{(.35)(.05)}{(.65)(.02) + (.35)(.05)} = \frac{.0175}{.0305} = .5738$$

Bayes' Theorem

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \ldots + P(A_n)P(B \mid A_n)}$$

Tabular Approach to Bayes' Theorem—2-Supplier Problem

(1) Events A_i	(2) Prior Probabilities $P(A_i)$	(3) Conditional Probabilities $P(B \mid A_1)$	(4) Joint Probabilities $P(A_i \cap B)$	(5) Posterior Probabilities $P(A_i \mid B)$
A_1	.65	.02	.0130	.0130/.0305 =.4262
A_2	.35	.05	.0175	.0175/.0305 =.5738
	1.00		P(B)=.0305	**1.0000**

Example

A simple example involving two possible prior events, B_1 and B_2, will illustrate the use of Bayes' theorem. A fragment of a hitherto unknown species of mosasaur has been found in a stream bed in western Kansas, and a vertebrate paleontologist would like to send a student field party out to search for more complete remains. Unfortunately, the source of the fragment cannot be identified with certainty because the fossil was found below the junction of two dry stream tributaries. The drainage basin of the larger stream contains about 18 mi^2, while the basin drained by the smaller stream includes only about 10 mi^2. On the basis of just this information alone, we might postulate that the probability that the fragment came from one of the drainage basins is proportional to the area of the basin, or

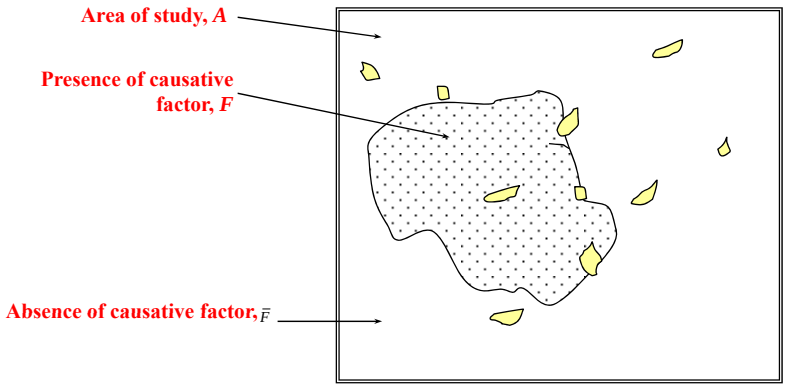
However, an examination of a geologic report and map of the region discloses the additional information that about 35% of the outcropping Cretaceous rocks in the larger basin are marine, while almost 80% of the outcropping Cretaceous rocks in the smaller basin are marine. We may therefore postulate the conditional probability that, given a fossil is derived from basin B_i, it will be a marine fossil, as proportional to the percentage of the Cretaceous outcrop area in the basin that is marine, or for basin B_1

$$p(A|B_1) = 0.35$$

and for basin B_2

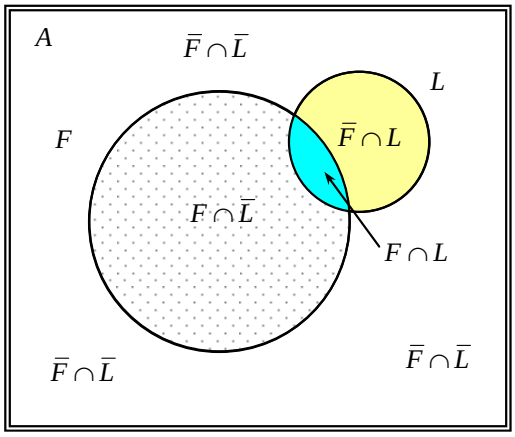
$$p(A|B_2) = 0.80$$

Landslide occurrence in an area



Venn diagram summarizing the spatial overlap relationships between the causative factor and the landslides.

Area in pixel
$N\{A\}$= 1000 (total area of study)
$N\{F\}$= 340 (area of causative factor)
$N\{L\}$= 110 (area of landslide)
$N\{F \cap L\}$= 70 (area of landslide on causative factor)



- Read paper
- Predictive modeling of rainfall-induced landslide hazard in the Lesser Himalaya of Nepal based on weights-of-evidence

Geomorphology 102 (2008) 496–510

Contents lists available at ScienceDirect

Geomorphology

journal homepage: www.elsevier.com/locate/geomorph

Predictive modelling of rainfall-induced landslide hazard in the Lesser Himalaya of Nepal based on weights-of-evidence

Ranjan Kumar Dahal [a,b,*], Shuichi Hasegawa [a], Atsuko Nonomura [a], Minoru Yamanaka [a], Santosh Dhakal [c], Pradeep Paudyal [d]

[a] Department of Safety Systems Construction Engineering, Faculty of Engineering, Kagawa University, 2217-20, Hayashi-cho, Takamatsu City, 761-0396, Japan
[b] Department of Geology, Tri-Chandra Multiple Campus, Tribhuvan University, Ghantaghar, Kathmandu, Nepal
[c] Department of Mines and Geology, Lainchaur, Kathmandu, Nepal
[d] Mountain Risk Engineering Unit, Tribhuvan University, Kirtipur, Kathmandu, Nepal

ARTICLE INFO

Article history:
Received 9 December 2007
Received in revised form 19 May 2008
Accepted 20 May 2008
Available online 4 June 2008

Keywords:
Lesser Himalaya
Nepal
Landslides
Weights-of-evidence
GIS
Landslide hazard mapping

ABSTRACT

Landslide hazard mapping is a fundamental tool for disaster management activities in mountainous terrains. The main purpose of this study is to evaluate the predictive power of weights-of-evidence modeling in landslide hazard assessment in the Lesser Himalaya of Nepal. The modelling was performed within a geographical information system (GIS), to derive a landslide hazard map of the south-western marginal hills of the Kathmandu Valley. Thematic maps representing various factors (e.g., slope, aspect, relief, flow accumulation, distance to drainage, soil depth, engineering soil type, landuse, geology, distance to road and extreme one-day rainfall) that are related to landslide activity were generated, using field data and GIS techniques, at a scale of 1:10,000. Landslide events of the 1978s, 1980s, and 1990s were used to assess the Bayesian probability of landslides in each unit with respect to the causative factors. To assess the accuracy of the resulting landslide hazard map, it was correlated with a map of landslides triggered by the 2002 extreme rainfall events. The accuracy of the map was evaluated by various techniques, including the area under the curve, success rate and prediction rate. The resulting landslide hazard values calculated from the old landslide data showed a prediction accuracy of >80%. The analysis suggests that geomorphological and human-related factors play significant roles in determining the probability value, while geological factors play only minor roles. Finally, after the restitution of the landslide hazard values of the new landslides using those of the old landslides, a landslide hazard map with >80% prediction accuracy was prepared. The methodology appears to have extensive applicability to the Lesser Himalaya of Nepal, with the limitation that the model's performance is contingent on the availability of data from past landslides.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Landslides are among the most damaging natural hazards in the mountainous terrains of the Lesser Himalaya of Nepal. Sites that are particularly at risk for landslides should therefore be identified so as to reduce damage in the region. Landslide hazard assessment has become a vital subject for authorities responsible for infrastructural development and environmental protection. Much research has been carried out to prepare landslide susceptibility and landslide hazard maps. According to Varnes (1984), landslide hazard in a given area can be assessed in terms of probability of occurrence of a potentially damaging landslide event within a specified period. Both intrinsic and extrinsic variables affect landslide hazards (Siddle et al., 1991; Wu and Sidle, 1995; Atkinson and Massari, 1998; Dai et al., 2001; Çevik and Topal, 2003). Intrinsic variables determining hazards include bedrock geology, topography, soil depth, soil type, slope gradient, slope aspect, slope curvature, elevation, engineering properties of the slope material, land use pattern, and drainage patterns. Extrinsic variables include heavy rainfall, earthquakes, and volcanic activities. Although the probability of landslide occurrence depends on both intrinsic and extrinsic variables, the latter possess a temporal distribution which is more difficult to handle in modelling practice. Therefore, for landslide hazard assessment, "landslide susceptibility mapping" is often conducted in which the extrinsic variables are not considered in determining the probability of landslide occurrence (Dai et al., 2001). In this research, a landslide hazard map was prepared by considering the extrinsic variable of rainfall in addition to the intrinsic variables.

There have been numerous studies involving landslide hazard evaluation (Guzzetti et al., 1999). Landslide hazard may be assessed through heuristic, deterministic, and statistical approaches (Yin and Yan, 1988; Van Westen and Terlien, 1996; Gökceoglu and Aksoy, 1996;

* Corresponding author. Department of Safety Systems Construction Engineering, Faculty of Engineering, Kagawa University, 2217-20, Hayashi-cho, Takamatsu City, 761-0396, Japan. Tel.: +81 87 864 2140; fax: +81 87 864 2038.
E-mail address: ranjan@dr.cje.net.jp, dahal@yahoo.com (R.K. Dahal).
URL: http://www.ranjan.net.np (R.K. Dahal).

0169-555X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.geomorph.2008.05.041

愛媛大学 EHIME UNIVERSITY

2nd Lecture

KAGAWA UNIVERSITY

Answer the following:

- What are prior odd and posterior odds?
- What is conditional or posterior probability of the landslide and how it can be expressed?
- What is Logits? Why authors used it.
- Why the odds of presence of landslide can be expressed as $O\{L\}=\frac{P\{L\}}{1-P\{L\}}$ and $O\{L\}=\frac{P\{L\}}{P\{\overline{L}\}}$?
- Why authors used logarithm of likelihood ratios?
- What do you mean by weights-of-evidence?

愛媛大学 EHIME UNIVERSITY

2nd Lecture

KAGAWA UNIVERSITY

Continuous Random Variables

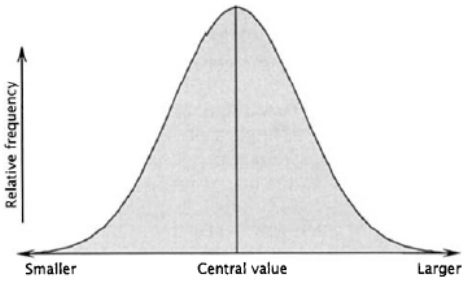
- In most experimental work, however, the possible outcomes are not discrete. Rather, there is an infinite continuum of possible results that might be obtained.
- The range of possible outcomes may be finite and in fact quite limited, but within the range the exact result that may appear cannot be predicted. Such events are called ***continuous random variables***.
- ***Suppose,*** we measure the length of the hinge line on a brachiopod and find it to be 6 mm long. By using a binocular microscope, a length of 6.2 mm, by using an optical comparator 6.23 mm, and with a scanning electron microscope, 6.231 mm.
- We can always find a difference between two measurements, if we conduct the measurements at a fine enough scale. The corollary of this statement is that every outcome on a continuous scale of measurement is unique, and that the probability of obtaining a specific, exact result must be zero!

Continuous Random Variables

- Permeability tests on core samples - always varied
- Unit weight of same type of soil in an are of 25 m^2 when measure in a grid of 1 m - always varied
- Cohesion of soil - always varied
- Variation induced into measurements by inaccuracy of instrumentation is most apparent when repeated measurements are made on a single object or a test is repeated without change. This variation is called ***experimental error.***

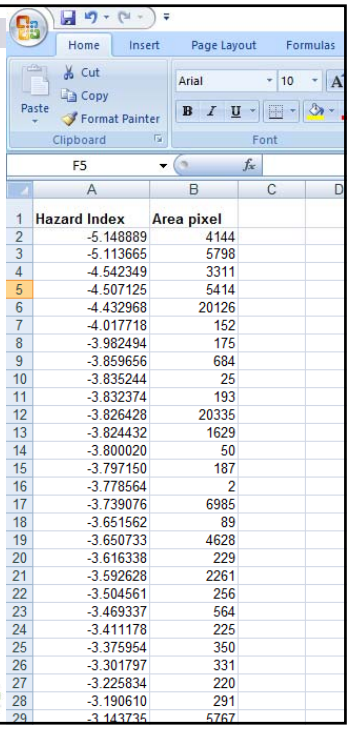
Normal distribution

- Repeated measurements on large samples drawn from natural populations may produce a characteristic frequency distribution. Most values are clustered around some central value, and the frequency of occurrence declines away from this central point.
- A graph of the distribution appears bell-shaped, and is called a *normal distribution. It often is assumed that random variables are normally* distributed, and many statistical tests are based on this supposition.



Do some statistical analysis

1. Calculate percentile area pixel.
2. Plot %cumulative frequency curve
3. Plot histogram of 10% interval.
4. Prepare *Box-and-whisker* plot.
5. Find mean, mode, standard deviation, coefficient of variation of hazard index in the study area (1 pixel = 20 m^2)
6. If we assume that the distribution of hazard index is normal, what will be the % ranges of hazard index value which will cover two-thirds of the hazard index.

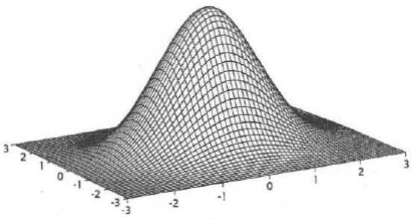


	A	B	C
1	Hazard Index	Area pixel	
2	-5.148889	4144	
3	-5.113665	5798	
4	-4.542349	3311	
5	-4.507125	5414	
6	-4.432968	20126	
7	-4.017718	152	
8	-3.982494	175	
9	-3.859656	684	
10	-3.835244	25	
11	-3.832374	193	
12	-3.826428	20335	
13	-3.824432	1629	
14	-3.800020	50	
15	-3.797150	187	
16	-3.778564	2	
17	-3.739076	6985	
18	-3.651562	89	
19	-3.650733	4628	
20	-3.616338	229	
21	-3.592628	2261	
22	-3.504561	256	
23	-3.469337	564	
24	-3.411178	225	
25	-3.375954	350	
26	-3.301797	331	
27	-3.225834	220	
28	-3.190610	291	
29	-3.143735	5767	

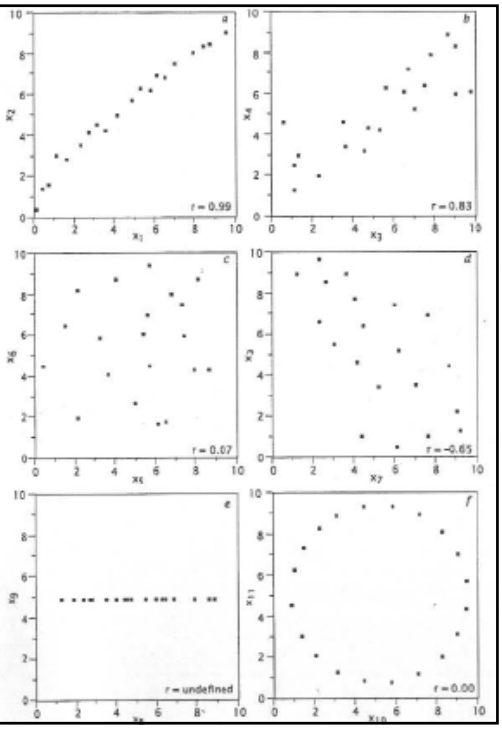
Joint Variation of Two Variables

- The variance of a single property can be extended to calculation of a measure of the mutual variability of a pair of properties.
- This measure called the ***covariance***, *is the joint variation of two variables about* their common mean.

Joint probability distribution of two independent normal distributions. Both x_1 *and* x_2 *are normally distributed.*



Scatter plot with difference covariance



Calculate Covariance

Table 2–3. Chromium, nickel, and vanadium in an Upper Pennsylvanian shale from Kansas.

	Cr (ppm)	Ni (ppm)	V (ppm)
	205	130	180
	255	165	215
	195	100	135
	220	135	200
	235	145	205
TOTALS =	1110	675	935
MEANS =	222	135	187

1. **Compute covariance between Cr and Ni, Ni and V, and Cr and V**
2. **Prepare two scatter diagrams of two variables with high covariance and low variance.**

Next class

- **Induced Correlations, Log ratio Transformation, Comparing Normal Populations, Central Limits Theorem**
- **Home work:**
  - Differentiate **"population" and "sample,"** with few geological data as examples?
  - Read paper of Ranjan and team
  - Do statistical analysis.
  - Calculate covariance and make scatter plot.
- **Homework Submission date Next week Friday (April 1st).**
- **Lecture notes in http://www.ranjan.net.np**