www.ranjan.net.np

Basic Statistics

Ranjan Kumar Dahal, PhD, PostDoc, M.ASCE

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

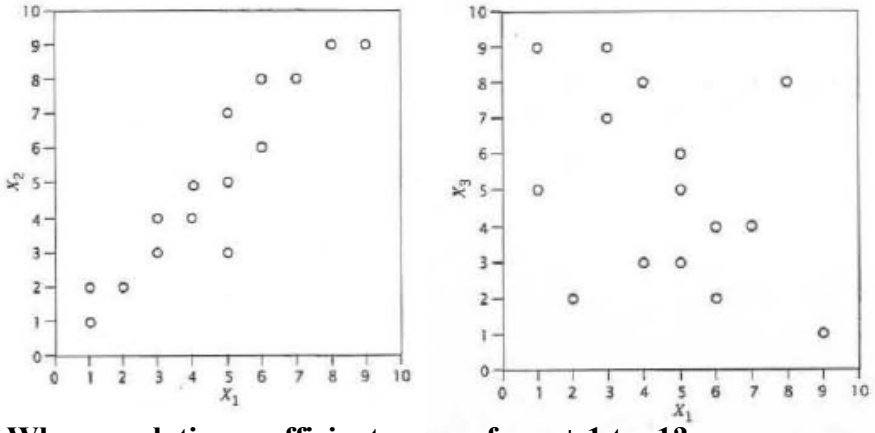
2nd Lecture

www.ranjan.net.np

Course Contents – first part

- 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.

Which one has high covariance?



Why correlation coefficient ranges from + 1 to -1?

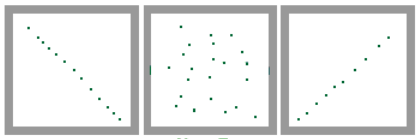
What do you mean by correlation between two variables *is* zero?

Are there guidelines to interpreting correlation coefficient?

	Coefficient, *r*	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

Covariance

- Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a *single* variable varies, **co** variance tells you how **two** variables vary together. Can be calculate by multiplying the correlation between the two variables by the standard deviation of each variable.
- **Advantages of the Correlation Coefficient**
  - The Correlation Coefficient has several advantages over covariance for determining strengths of relationships:
  - Covariance can take on practically any number while a correlation is limited: -1 to +1.
  - Because of it's numerical limitations, correlation is more useful for determining **how strong** the relationship is between the two variables.
  - Correlation does not have units. Covariance always has units
  - Correlation isn't affected by changes in the center (i.e. mean) or scale of the variables



Large Negative Covariance | Near Zero Covariance | Large Positive Covariance

Induced Correlations

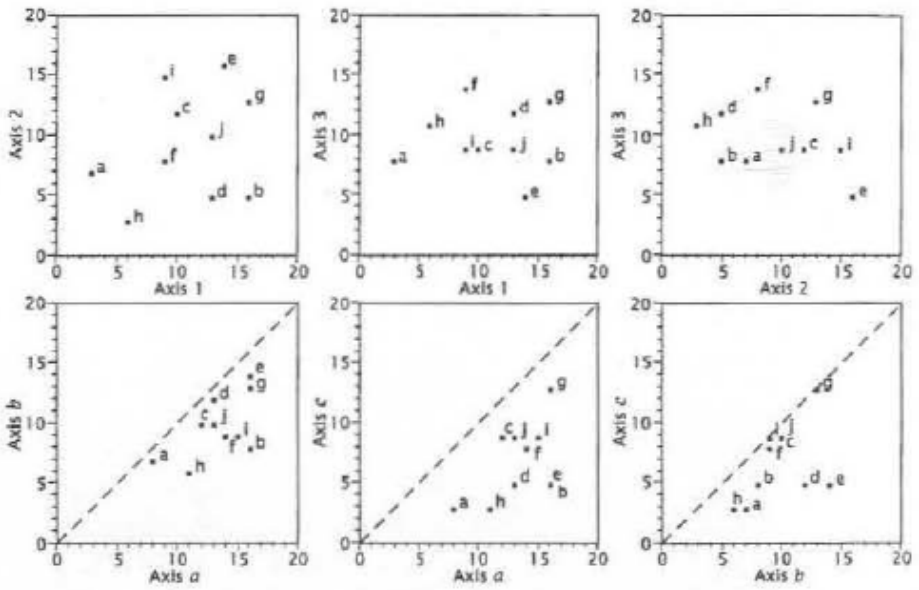
- **Some correlations between variables do not reflect the relationships between them, but are induced by an operation or transformation that has been performed on the variables. Such correlation is known as induced correlations**

Pebble	Axis 1	Axis 2	Axis 3
a	3	7	8
b	16	5	8
c	10	12	9
d	13	5	12
e	14	16	5
f	9	8	14
g	16	13	13
h	6	3	11
i	9	15	9
j	13	10	9
TOTALS	109	94	98
MEANS	10.9	9.4	9.8
CORRELATIONS	$r_{1,2} = 0.279$	$r_{1,3} = -0.021$	$r_{2,3} = -0.349$

Pebble	*a*-Axis	*b*-Axis	*c*-Axis
a	8	7	3
b	16	8	5
c	12	10	9
d	13	12	5
e	16	14	5
f	14	9	8
g	16	13	13
h	11	6	3
i	15	9	9
j	13	10	9
TOTALS	134	98	69
MEANS	13.4	9.8	6.9
CORRELATIONS	$r_{a,b} = 0.597$	$r_{a,c} = 0.499$	$r_{b,c} = 0.467$

Randomly selected and We might suppose that the measurements will be correlated, because a large pebble will most likely have large values for all three axes and, conversely, a small pebble will have small measurements for all three axes.

愛媛大学 EHIME UNIVERSITY　KAGAWA UNIVERSITY



Scatter diagrams of axial length, in cm, of pebbles collected on a shingle beach. Top: Original measurements collected in random order. Bottom: Measurements sorted into *a-, b-.* and c-axes, which causes all points to plot below diagonals of diagrams.

Logratio Transformation

- **Aitchison (1986) provides an extensive treatment of closure and offers a number of transformations that convert compositional variables into forms that can be analyzed by conventional statistical techniques.**

logratio variances

$$s^2_{j/k} = \text{var}\left(\ln \frac{x_j}{x_k}\right)$$

Table 2–10. Chemical compositions, in oxide percents, of 11 pyroxenes. R_2O_3 is the sum of Al_2O_3, TiO_2, Fe_2O_3, and Cr_2O_3.

SiO_2	R_2O_3	FeO	MgO	CaO
57.75	1.87	3.65	36.50	0.23
49.97	4.20	28.64	15.75	1.44
46.48	1.33	47.21	3.53	1.45
54.99	4.70	1.59	17.27	21.45
48.72	1.89	26.85	1.07	21.47
52.85	5.02	5.71	16.48	19.94
47.58	5.67	21.73	7.42	17.60
50.11	4.67	18.90	16.33	9.99
49.79	8.79	26.33	6.99	8.10
53.06	1.73	17.52	23.61	4.08
50.51	3.53	29.21	12.89	3.86

logratio variances

$$s^2_{j/k} = \mathrm{var}\left(\ln \frac{x_j}{x_k}\right)$$

Table 2–11. Ten unique logratio transforms of pyroxene compositional variables listed in Table 2–10.

$\frac{SiO_2}{R_2O_3}$	$\frac{SiO_2}{FeO}$	$\frac{SiO_2}{MgO}$	$\frac{SiO_2}{CaO}$	$\frac{R_2O_3}{FeO}$	$\frac{R_2O_3}{MgO}$	$\frac{R_2O_3}{CaO}$	$\frac{FeO}{MgO}$	$\frac{FeO}{CaO}$	$\frac{MgO}{CaO}$
3.430	2.761	0.459	5.526	−0.669	−2.971	2.096	−2.306	2.764	5.067
2.476	0.557	1.155	3.547	−1.920	−1.322	1.070	0.598	2.990	2.392
3.554	−0.016	2.578	3.467	−3.569	−0.976	−0.086	2.593	3.483	0.890
2.460	3.543	1.158	0.941	1.084	−1.301	−1.518	−2.385	−2.602	−0.217
3.250	0.596	3.818	0.819	−2.654	0.569	−2.430	3.223	0.224	−2.999
2.354	2.225	1.165	0.975	−0.129	−1.189	−1.379	−1.060	−1.251	−0.191
2.127	0.784	1.858	0.995	−1.344	−0.269	−1.133	1.0745	0.211	−0.864
2.373	0.975	1.121	1.613	−1.398	−1.252	−0.760	0.146	0.636	0.491
1.734	0.637	1.963	1.816	−1.097	0.229	0.082	1.326	1.179	−0.147
3.423	1.108	0.810	2.565	−2.315	−2.614	−0.858	−0.298	1.457	1.755
2.661	0.548	1.366	2.572	−2.113	−1.295	−0.089	0.818	2.024	1.206

Table 2–12. Compositional variation array of logratio-transformed variables in Table 2–11. Upper diagonal contains variances, lower diagonal contains means.

	SiO_2	R_2O_3	FeO	MgO	CaO
SiO_2		0.3682	1.2187	0.8868	2.1653
R_2O_3	2.7129		1.6262	1.1202	1.5991
FeO	1.2471	−1.4658		3.2009	3.3923
MgO	1.5865	−1.1264	0.3393		4.1854
CaO	2.2578	−0.4552	1.0106	0.6713	

- **Aitchison (1986) shows that the logratio covariances also can be calculated from the logratio variances**

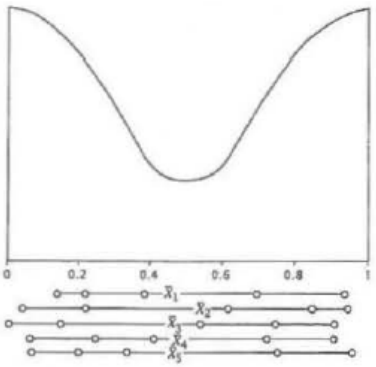
- **Aitchison proposes the use of the *centered logratio covariance* *in* which the divisors of the logratios are the geometric means of the original compositional variables.**

Table 2–14. Centered logratio covariances for pyroxene analyses.

	$\ln\left(\frac{SiO_2}{g}\right)$	$\ln\left(\frac{R_2O_3}{g}\right)$	$\ln\left(\frac{FeO}{g}\right)$	$\ln\left(\frac{MgO}{g}\right)$	$\ln\left(\frac{CaO}{g}\right)$
$\ln(SiO_2/g)$	0.00409	−0.00174	−0.05691	0.04619	−0.02665
$\ln(R_2O_3/g)$	−0.00174	0.36066	−0.08234	0.10777	0.43474
$\ln(FeO/g)$	−0.05691	−0.08234	1.10086	−0.56250	−0.09180
$\ln(MgO/g)$	0.04619	0.10777	−0.56250	0.97505	−0.55122
$\ln(CaO/g)$	−0.02665	0.43474	−0.09180	−0.55122	2.10791

Aitchison's transformation removes spurious negative correlations between compositional variables, but in some circumstances there are difficulties 1n its application. Because the transformation uses logarithms, it cannot be applied if some are zero. So need to assign small amount near to zero.


- **Comparing Normal Populations – self study**
- ***Central limits theorem* states that if sets of random samples are taken from any population, and the means calculated for these samples, the sample means will tend to be normally distributed.**
- **The tendency toward normality becomes more pronounced for samples of larger size.**




Testing the Mean

- The comparison between the difference in means and the standard error can be made in the this way
- The test statistic, *z,* is normally distributed with a mean of zero and a standard deviation of one, if the sample mean was indeed drawn from the hypothesized population.
- If *z is* excessively large, we will tend to conclude that the sample was not taken from this population.

$$z = \frac{\overline{X} - \mu}{s_e} = \frac{\overline{X} - \mu}{\sigma\sqrt{1/n}}$$

null hypothesis

$$H_0 : \mu_1 = \mu_2$$

Alternative hypothesis

$$H_1 : \mu_1 \neq \mu_2$$

Stating that the mean of the population from which the sample was drawn does not equal the specified population mean.

This combination produces four possible outcomes, of which two are correct and two incorrect.

	Hypothesis is correct	Hypothesis is incorrect
Hypothesis accepted	Correct decision	Type II error, β
Hypothesis rejected	Type I error, α	Correct decision

Either acceptance of a true hypothesis or rejection of a false hypothesis will result in a correct decision. If a null hypothesis is rejected when it is in fact true, a type I error has been committed. Conversely, if an erroneous hypothesis is accepted, a type II error is committed. In terms of our example, the illustration above may be redrawn:

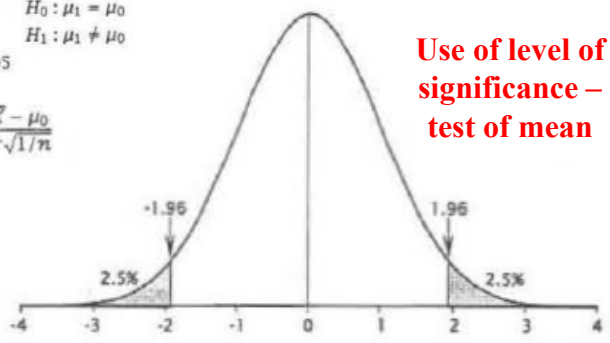
	Actuality	
Hypothesis	Slab is *Composita*	Slab is not *Composita*
μ of slab = μ of *Composita*	Correct decision	Type II error, β
μ of slab ≠ μ of *Composita*	Type I error, α	Correct decision

Level of significance

- In standard statistical procedures, the probability of committing a type I error is called the ***level of significance*** ***and is denoted by α; this probability must be*** specified before running the test.
- In order to minimize the possibility of committing a type II error, we express the null hypothesis with the intention that it will be rejected.

1. The hypothesis and alternative: $H_0 : \mu_1 = \mu_0$, $H_1 : \mu_1 \neq \mu_0$
2. The level of significance: $\alpha = 0.05$
3. The test statistic: $z = \dfrac{\overline{X} - \mu_0}{\sigma\sqrt{1/n}}$

1. $H_0 : \mu_{slab} = 14.2$ mm, $H_1 : \mu_{slab} \neq 14.2$ mm
2. $\alpha = 0.05$
3. $z = \dfrac{20.0 - 14.2}{4.7\sqrt{1/6}} = 3.023$



Use of level of significance – test of mean

if z falls into the right-hand region, the mean of the sample's parent population is larger than the mean of the known population. From Appendix Table A.1, we find that approximately 2.5% of the area of the curve is to the left of a ***critical value*** of $z = -1.9$, and 97.5% (100% − 2.5% = 97.5%) is to the left of a critical value of $z = +1.9$. The computed test value of $z = 3.0$ exceeds 1.9, so we conclude that the means of the two populations are not equal, and the collection of fossils must represent some genus other than *Composita*. It is important to note the assumptions that have been made in the application of this test. The normal test assumes:

1. The sample of brachiopods was selected randomly.
2. The population of lengths of *Composita* is known to be normally distributed.
3. The variance in lengths of *Composita* is known to be 22.1 mm.

Self study

- P-Values
- Significance
- *Confidence interval*
- The t-Distribution
- Degree of freedom
- Confidence intervals based on *t*
- A test of the equality of two sample means
- The *t -test of correlation*
- The F-Distribution

Next class

- Analysis of variance
- Home work:
  - What do you mean by closed data set and open data set? Write drawbacks of closed data set in induced correlation.
  - Why null hypothesis and its alternative are mutually exclusive and all inclusive?
  - What is one tailed test? Use Table 2.16 and perform calculation.
  - Define "Confidence intervals based on t" with example.
  - Read paper "DEM-based deterministic landslide hazard analysis in the Lesser Himalaya of Nepal" answer how can error propagation be done when we used uniformly used random variables?
  - In F-Distribution, what will be interpretation of data if "Null hypotheses cannot be rejected"
  - Download data from http://www.kgs.ku.edu/Mathgeo/Books/Stat/Index.html and do exercise 2.1, 2.2, and 2.3.
  - Home work submission: May 29, 2017, 10 PM.
- Lecture notes in http://www.ranjan.net.np