University of Colombo
Faculty of Arts
Postgraduate Diploma in Economic Development (DED) – 2016
DED 504: Applied Social Statistics
Answer Any Four (04) Questions Only
Time Allowed: 03 Hours Only
Use of Any Type of Calculators is Allowed

---

1) The lecturers at a university are required to submit their final examination paper to the Exams and Timetabling office 10 days before the end of teaching for that semester. The exam coordinator sampled 20 lectures and recorded the number if days before the final exam that each submitted his or her exam. The results are:

12	0	10	4	13	12	9	11	4	8
6	5	2	13	3	9	8	4	14	7

a) Calculate the three statistics that measure central location.
(03 Marks)

b) Briefly describe what each statistic in part (a) tells you.
(06 Marks)

c) What are the limitations of mean, median and mode?
(08 Marks)

d) Calculate Coefficient of Variance (CV) and interpret them.
(08 Marks)

2) A statistics practitioner took a random sample of 50 observations from a population whose standard deviation is 25 and computed the sample mean to be 100.

a) Estimate the population mean with 90% confidence.
(05 Marks)

b) Repeat part (a) using a 95% confidence level.
(05 Marks)

c) Repeat part (a) using a 99% confidence level.
(05 Marks)

d) Describe the effect on the confidence interval estimate of increasing the confidence level.
(10 Marks)

3) The heights of children two years old are normally distributed with a mean of 80cm and a standard deviation of 3.6cm. Paediatricians regularly measure the heights of toddlers to determine whether there is a problem. There may be a problem when a child is in the top or bottom 5% of heights.

a) Determine the heights of two-year old children that could be a problem.
(10 Marks)

b) Find the probability of these events

i) A two-year old child is taller than 90 cm

(04 Marks)

ii) A two-year old child is shorter than 85 cm

(04 Marks)

iii) A two-year old child is between 75 and 85 cm

(07 Marks)

4) Late payment of medical claims can add to the cost of health care. An article (M. Freudenheim, "The Check Is Not in the Mail," *The New York Times*, May 25, 2006, pp.C1 C6) reported that the mean time from the date of service to the date of payment for one insurance company was 41.4 days during a recent period. Suppose that a sample of 100 medical claims is selected during the latest time period. The sample mean time from the date of service to the date of payment was 39.6 days, and the sample standard deviation was 7.4 days.

a) Using the 0.05 level of significance, is there evidence that the population mean has changed from 41.4 days?

(10 Marks)

b) What is your answer in (a) if you use 0.01 level of significance?

(8 Marks)

c) What is your answer in (a) if the sample mean is 38.2 days and the sample standard deviation is 10.7 days?

(7 Marks)

5) A large corporation is interested in determining whether a relationship exists between the commuting time of its employees and the level of stress-related problems observed on the job. A study of 116 assembly-line works reveals the following information. At the 0.01 level of significance, is there any evidence of a significant relationship between commuting time and stress level?

Commuting Time	Stress Level			Total
	High	Moderate	Low	
Under 15 minutes	9	5	18	32
15-45 minutes	17	8	28	53
Over 45 minutes	18	6	7	31
Total	44	19	53	116

a) State null and alternative hypotheses of this test.

(04 Marks)

b) Construct frequency distribution if null hypothesis is true.

(04 Marks)

c) Determine the critical value for this test using CHI-squared distribution with relevant degree of freedom.

(05 Marks)

d) Calculate the test statistic.

(06 Marks)

e) State the conclusion.

(06 Marks)

6) A dean of an art faculty waned to determine whether or not there were differences in annual salaries two years after graduation among those who majored in geography, sociology and economics. As a preliminary experiment, he surveyed six graduates from each of the three subject areas and asked how much they earned annually. The results in EXCEL output is given below with some missing information. Do these data provided sufficient evidence to allow the dean to conclude that there are differences between the majors' incomes?

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Geography	6	222	37.0	70.4
Sociology	6	201	33.5	60.3
Economics	6	174	29.0	33.2

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	?	?	?	?	0.2047
Within Groups	?	?	54.633		
Total	1012.5	?			

a) State null and alternative hypotheses of this test.
(04 Marks)

b) Fill the ANOVA table with missing information.
(06 Marks)

c) Find the critical value from F distribution at 5% level of significance.
(05 Marks)

d) Test the hypothesis at 5% level of significance and interpret the results.
(10 Marks)

7) An economist wanted to investigate the relationship between office rents and vacancy rates. Accordingly, he took a random sample of monthly office rents and the percentage of vacant office space in 30 different cities.

Variable	Coefficients	Standard Error	t Stata	Prob.
Constant	20.63971	1.14279	18.0608	0.0000
VACANCY	-0.3038	0.08958	-3.3912	0.0021

R-Squared: 29.11%
Observations 30

a) Write the regression equation
(04 Marks)

b) Interpret the coefficients.
(08 Marks)

c) Can we conclude at the 5% significance level that higher vacancy rates result in lower rents?

(08 Marks)

d) Measure how well the linear model fits the data. Discuss what this measure tells you.

(05 Marks)

8) Write brief explanatory notes on any five (05) (05 marks each)

a) Type 1 and Type 2 errors
b) Interquartile rage
c) Skewed distribution
d) ANOVA Assumptions
e) Two-factor analysis
f) Four components of time series analysis
g) Coefficient of determination
h) Assumptions of regression analysis
i) Measures of forecast accuracy
j) Method of exponential smoothing

Measure of Central Location and Dispesion

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}{N} \qquad s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

$$\text{CV} = \frac{\text{Standard Devition}}{\text{Mean}} \times 100\% \qquad L_p = (N+1) \times \frac{P}{100} \qquad L_p = (n+1) \times \frac{p}{100}$$

Normal Distribution

$X \sim \text{Normal}(\mu; \sigma^2)$ $\qquad Z = \frac{X-\mu}{\sigma}$

Sample Distribution

$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$

Sampling Distribution for Sample Proportion

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}} \qquad z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$$

Confidence Intervals

Quantitative or Numerical data:

$$\bar{x} \pm z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \bar{x} \pm t_{\alpha/2,n-1}\, s/\sqrt{n} \qquad n = \left[\frac{z_{\alpha/2}\sigma}{B}\right]^2 \text{ or } \left[\frac{z_{\alpha/2}^2\sigma^2}{B^2}\right]$$

Qualitative or Categorical data:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ or } \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \qquad n = \left[\frac{z_{\alpha/2}\sqrt{\hat{p}\hat{q}}}{B}\right] \text{ or } \left[\frac{z_{\alpha/2}^2\hat{p}\hat{q}}{B^2}\right] \text{ where } \hat{q} = 1 - \hat{p}$$

Hypothesis Testing

Quantitative or Numerical data:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Qualitative or Categorical data:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \text{ where } q = 1 - p$$

Chi-squared tests

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

Single-factor analysis (one way analysis of variance)

$$SST = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 \quad \text{or} \quad SSE = \sum_{j=1}^{k}(n_j - 1)s_j^2 \quad \text{where} \quad s_j^2 = \frac{1}{(n_j - 1)}\left[\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2\right]$$

$$MST = \frac{SST}{k-1} \qquad MSE = \frac{SSE}{n-k} \qquad F = \frac{MST}{MSE}$$

Two-factor analysis (Two way analysis of variance)

$$SS(Total) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk} - \bar{\bar{x}})^2 \qquad SS(A) = rb\sum_{i=1}^{a}(\bar{x}[A]_i - \bar{\bar{x}})^2 \qquad SS(B) = ra\sum_{i=1}^{b}(\bar{x}[B]_i - \bar{\bar{x}})^2$$

$$SS(AB) = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}[AB]_{ij} - \bar{x}[A]_i - \bar{x}[B]_j + \bar{\bar{x}})^2 \qquad SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk} - \bar{x}[AB]_{ij})^2$$

$$MS(A) = \frac{SS(A)}{a-1} \qquad MS(B) = \frac{SS(B)}{b-1} \qquad MS(AB) = \frac{SS(AB)}{[(a-1)(b-1)]} \qquad MSE = \frac{SSE}{n-ab}$$

$$F_A = \frac{MS(A)}{MSE} \qquad F_B = \frac{MS(B)}{MSE} \qquad F_{AB} = \frac{MS(AB)}{MSE}$$

Correlation Analysis and Simple Linear regression Analysis

$$SS_x = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = \sum x_i^2 - n\bar{x}^2$$

$$SS_y = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = \sum y_i^2 - n\bar{y}^2$$

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = \sum x_i y_i - n\bar{x}\bar{y}$$

Correlation Coefficient

$$r = \frac{SS_{xy}}{\sqrt{(SS_x)(SS_y)}} \qquad t_{n-2} = \frac{r - p}{\sqrt{\frac{1-r^2}{n-2}}}$$

Sample regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \hat{\beta}_1 = \frac{SS_{xy}}{SS_x} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

SSE and Standard error of estimate

$$SSE = SS_y - \frac{SS_{xy}^2}{SS_x} \qquad s_\varepsilon^2 = \frac{SSE}{n-2} \quad \text{or} \quad s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

Test statistics for the significance of the population slope

$$t_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \qquad s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{SS_x}}$$

Coefficient of Determination

$$R^2 = \frac{SS_{xy}^2}{(SS_x)(SS_y)} = \frac{SSR}{SS_y} = 1 - \frac{SSE}{SS_y} = 1 - \frac{SSE}{SST} \qquad R^2 = r^2$$