MATH 545
DR. MCLOUGHLIN'S CLASS
STATISTICAL FORMULAE FOR CORRELATION
HANDOUT VIII

Recall we are attempting to estimate parameters in a population (e.g.: a population mean μ,

a population variance σ^2, a population standard deviation σ, or any other parameter let us call it θ) so,

Let $D = \{X_1, X_2, X_3, \ldots, X_n\}$ be a finite data set from a population of interest.

Let $X_1, X_2, X_3, \ldots, X_n$ be the finite random sample.

Recall these statistical formulae from previous handouts:

$$\overline{X} = \frac{\sum_{k=1}^{n} X_k}{n} \quad \text{and} \quad \overline{X} \text{ is } \hat{\mu}$$

$$S^2 = \frac{\sum_{k=1}^{n} (X_k - \overline{X})^2}{n-1} \quad \text{and} \quad S^2 \text{ is } \hat{\sigma}^2$$

$$s = \sqrt{= \frac{\sum_{k=1}^{n} (X_k - \overline{X})^2}{(n-1)}} \quad \text{and} \quad S \text{ is } \hat{\sigma}$$

<u>Definition 1</u>: If X and Y are random variables, and the function given by *f*(x, y) for each x and y in the domain of the function is the p. d. f. or p. m. f. at x and y. Then

the covariance of X and Y is $Cov[X, Y] = E[(X - \mu_x)(Y - \mu_y)]$

Note: We write $Cov[X, Y] = \sigma_{xy} = \sigma_{yx}$ and since $Cov[X, Y] = \sigma_{xy} = \sigma_{yx}$ it is generally the case in a bivariate discussion to denote $Var[X] = \sigma_{xx}$ and $Var[Y] = \sigma_{yy}$.

<u>Theorem 1</u>: If X and Y are random variables, and the function given by *f*(x, y) for each x and y in the domain of the function is the p. d. f. or p. m. f. at x and y. Then

$Cov[X, Y] = E[XY] - E[X]E[Y] = \mu_{xy} - (\mu_X \cdot \mu_Y)$

Definition 2: If X and Y are random variables, and the function given by *f*(x, y) for each x and y in the domain of the function is the p. d. f. or p. m. f. at x and y. Then

the Pearson product-moment correlation coefficient (or just correlation) of X and Y is $\rho[X, Y] = \frac{Cov[X, Y]}{\sigma_x \sigma_y}$. We also write $Corr[X, Y] = \rho_{xy}$

Theorem 2: If X and Y are random variables, and the function given by *f*(x, y) for each x and y in the domain of the function is the p. d. f. or p. m. f. at x and y. Then

$\sigma_{XY} \in [0, \infty)$

Theorem 3: If X and Y are random variables, and the function given by *f*(x, y) for each x and y in the domain of the function is the p. d. f. or p. m. f. at x and y. Then

$\rho_{XY} \in [-1, 1]$

Now suppose we have paired data such that

Let $D_1 = \{X_1, X_2, X_3, \ldots, X_n\}$ be a finite data set from a population of interest.

Let $D_2 = \{Y_1, Y_2, Y_3, \ldots, Y_n\}$ be a finite data set from a population of interest.

(either X_i, and Y_i are two measures of an attribute of a subject or could be paired because of some justification in the research area of the researcher.

The sample Pearson product-moment correlation is

$$r_{XY} = \frac{\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y})}{\sqrt{\sum_{k=1}^{n}(X_k - \overline{X})^2 \sum_{k=1}^{n}(Y_k - \overline{Y})^2}} = \frac{\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y})}{n\sqrt{(S_x)(S_y)}} = \frac{\sum_{k=1}^{n}(z_{X_k})(z_{Y_k})}{n}$$

r_{XY} is $\hat{\rho}_{XY}$

$r_{XY} \in [-1, 1]$

The sample Pearson product-moment correlation is a measure of the linear association between two variables. It is **not** a measure of causation, it does not show X creates Y, X causes Y, Y creates X, Y causes X, etc.

The heuristic for correlation is:

'low' near zero, 'moderate' not near zero nor near – 1 nor 1, 'high' near – 1 or 1.

Two variables X and Y can have 'low' correlation ('near zero') and still be associated.

Scatterplots assist us in noting correlations.

1. Find r_{XY} (if it exists) for the data sets; if it doesn't state why it doesn't:

A.

X	Y
0	0
1	1
2	2
3	3
4	4

B.

X	Y
0	1
1	3
2	5
3	7
4	9

C.

X	Y
0	8
1	6
2	3
3	0
4	-5

D.

X	Y
3	
5	8
	5
8	7
9	9

E.

X	Y
0	-1
1	0
0	1
-1	0
$\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{2}}{2}$
$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$
$-\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$
$-\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{2}}{2}$

T-test for Correlated Data Formula

Let $D_1 = \{X_1, X_2, X_3, \ldots, X_{n_1}\}$ be a finite data set from a population of interest.

Let $D_2 = \{Y_1, Y_2, Y_3, \ldots, Y_{n_2}\}$ be a finite data set from a population of interest.

Recall the sample Pearson product-moment correlation is

$$r_{XY} = \frac{\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y})}{\sqrt{\sum_{k=1}^{n}(X_k - \overline{X})^2 \sum_{k=1}^{n}(Y_k - \overline{Y})^2}} = \frac{\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y})}{n\sqrt{(S_x)(S_y)}} = \frac{\sum_{k=1}^{n}(z_{X_k})(z_{Y_k})}{n}$$

If the samples are related (two measures from the same subject or matched pairs), a correlated data formula is used (and let $n_1 = n_2$):

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2} - 2r_{XY}\left(\frac{S_X}{\sqrt{n_1}} \cdot \frac{S_Y}{\sqrt{n_2}}\right)}}$$