

Tammy M. Urban  
STAT 8240

Homework #1  
June 16, 2008

1. The procedure used for stratified sampling is **PROC SURVEYSELECT** with the option **STRATA**. The data must be sorted by the strata. The sample size (N) is the number of subjects/observations from each strata.

The data are various scores and demographic information regarding the 91 Algebra I students I taught during spring semester 2008. This dataset consists of the following variables.

VARIABLE	Description	Type	Values
PD	class period in which students were enrolled	ordinal	2, 3, or 4
MTM	midterm score	quantitative	
SMAvg	skills maintenance quiz average	quantitative	
PracEOCT	practice EOCT score	quantitative	
TESTAvg	semester test average	quantitative	
Missing	# of missing (not made up) assignments	ordinal	0-20
Zeros	# of zeros (failed to complete)	ordinal	0-10
Absences	# of class absences	ordinal	0-20
PT1Avg	average in part 1 of course (1st semester)	quantitative	
P1Teach	part 1 teacher	nominal	names
Race	race	nominal	W=white, B=black, H=Hispanic, A=Asian, M=Multiracial
Gender	gender	nominal	1=female, 2=male
CRCT_8	8th grade CRCT	ordinal	1=DNM, 2=MET, 3=EXC
DNM_MS	# of "does not meet" scores in Middle School	ordinal	0-3
EXC_MS	# of "exceeds" scores in Middle School	ordinal	0-3
ED	economically disadvantaged	ordinal	0=no, 1=yes
Class	class standing	ordinal	9=freshman, 10=sophomore, 11=junior, 12=senior
EOCT	Algebra I EOCT score	quantitative	

```

****      Import Spring EOCT Data      ****;

libname mysas 'T:\KSUMSAS';

PROC IMPORT OUT=MYSAS.Spring08
  DATAFILE='T:\KSUMSAS\spring08.xls'
  DBMS=EXCEL REPLACE;
  GETNAMES=YES;
RUN;

****      Sort Spring EOCT Data      ****;

DATA SORTED;
  SET MYSAS.SPRING08;
PROC SORT;
  by Gender PD;
RUN;

```



## 2. Exercise 17, page 92

- (a) The interval in terms of  $x$  is  $[0, \infty)$
- (b)  $y = \sqrt{x}$  or  $y^2 = x$  if a “non-function” is permitted.

## 3. Exercise 12, page 91

There is natural variation in nature and processes and is what makes things interesting. This variation is often called “noise”. While ideally, we would like no noise, this is not realistic.

When noise (variation) is extreme and markedly different from the rest of the data, we call these outliers. Not all noise is considered an outlier and some outliers may not even be included in the noise pattern if they are very extreme.

Outliers are not desirable as they can indicate one of several potential problems with data collection and/or coding.

In accommodating too much noise (variation) it is possible for an atypical value to be considered typical. This could be problematic. Conversely, if the noise results in shifting the pattern in some way, it is possible for a typical value to be considered atypical.

## 4. Exercise 2, page 89

- (a) Binary – ordinal: there is some inherent order in AM and PM
- (b) Continuous – ratio: assuming the light meter is measuring lumens or some other quantitative value of brightness that can be divided into smaller measures.
- (c) Discrete – ordinal: there would be an inherent order in most peoples’ classification of brightness.
- (d) Continuous – ratio: angle measures can be divided into smaller, fractional measures.
- (e) Discrete – ordinal: there is a clear order in the medal types
- (f) Continuous – ratio: heights can be divided into smaller, fractional measures.
- (g) Continuous – interval: even if a hospital contained an “infinite” number of patients, there is no such thing as  $\frac{1}{2}$  or  $\frac{1}{4}$  of a patient.
- (h) Discrete – nominal: groups of numbers categorize book publishing information with no inherent order in these categorizations.
- (i) Discrete – ordinal: there is an inherent order in how much light passes through but these are categorized.
- (j) Discrete – ordinal: there is an inherent order in military rank.
- (k) Continuous – ratio: distance can be divided into smaller, fractional measures.
- (l) Continuous – ratio: density measurements can be divided into smaller, fractional measures
- (m) Continuous – interval: these numbers are probably on a “reel” of many thousands of numbers but there are no decimals.

5. Exercise 18, page 92

(a) **Hamming distance:**

$$\text{dist}(x, y) = \left( \sum_{k=1}^n |x_k - y_k|^r \right)^{\frac{1}{r}} \quad \text{where } n = 10, r = 1$$

$$\text{dist}(x, y) = \sum_{k=1}^{10} |x_k - y_k| = 3$$

$$\begin{array}{rcccccccccc} x & = & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ y & = & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ \hline x - y & = & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 \\ |x - y| & = & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}$$

**Jaccard similarity:**

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \quad M_{01} = 1, M_{10} = 2, M_{11} = 2$$

$$\begin{array}{rcccccccccc} x & = & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ y & = & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ \hline & & & M_{11} & & M_{10} & & M_{11} & M_{01} & & & M_{10} \end{array}$$

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{2}{1 + 2 + 2} = \frac{2}{5} = 0.4$$

- (b) The Jaccard Similarity is more similar to the Simple Matching Coefficient because both of these are ratios of common situations.

The Hamming distance is more similar to the Cosine Similarity because both of these are a power of a sum of values.

- (c) The Jaccard Similarity should be used for comparing the number of genes shared by two organisms from different species because similarity compares what they share when they probably have more things NOT in common, whereas the Hamming distance would compare differences when they have more things in common.
- (d) The Hamming distance should be used for comparing two organisms of the same species because distances compare differences when they have more things in common and similarities compare what they share when they have more things NOT in common.

## 6. Exercise 19, page 93

(a)

$$\begin{array}{r} x = 1 \quad 1 \quad 1 \quad 1 \\ y = 2 \quad 2 \quad 2 \quad 2 \end{array}$$

### Cosine:

$$\cos(x, y) = (x \cdot y) / \|x\| \|y\| = 8 / (2 \cdot 4) = 1$$

$$(x \cdot y) = 1 \cdot 2 + 1 \cdot 2 + 1 \cdot 2 + 1 \cdot 2 = 8$$

$$\|x\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = \sqrt{4} = 2$$

$$\|y\| = \sqrt{2^2 + 2^2 + 2^2 + 2^2} = \sqrt{16} = 4$$

### Correlation:

$$\text{Corr}(x, y) = \frac{x' \cdot y'}{n-1} \quad \text{where } x' = \frac{x - \bar{x}}{\text{std}(x)} \quad \text{and} \quad y' = \frac{y - \bar{y}}{\text{std}(y)}$$

Correlation is undefined for these vectors as the standard deviations for both data sets is 0 – division by zero undefined.

### Euclidean:

$$\text{dist}(x, y) = \left( \sum_{k=1}^n |x_k - y_k|^r \right)^{\frac{1}{r}} \quad \text{where } n = 4, r = 2$$

$$\text{dist}(x, y) = \left( \sum_{k=1}^4 |x_k - y_k|^2 \right)^{\frac{1}{2}} = xxx$$

$$\sum_{k=1}^4 |x_k - y_k|^2 = |1-2|^2 + |1-2|^2 + |1-2|^2 + |1-2|^2 = 4$$

$$\left( \sum_{k=1}^4 |x_k - y_k|^2 \right)^{\frac{1}{2}} = 4^{\frac{1}{2}} = 2$$

(b)

$$\begin{aligned}x &= 0 & 1 & 0 & 1 \\y &= 1 & 0 & 1 & 0\end{aligned}$$

**Cosine:**

$$\cos(x, y) = (x \cdot y) / \|x\| \|y\| = 0 / 2 = 0$$

$$(x \cdot y) = 0 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 1 \cdot 0 = 0$$

$$\|x\| = \sqrt{0^2 + 1^2 + 0^2 + 1^2} = \sqrt{2}, \quad \|y\| = \sqrt{1^2 + 0^2 + 1^2 + 0^2} = \sqrt{2}$$

**Correlation:**

$$\text{Corr}(x, y) = \frac{x' \cdot y'}{n-1} \quad \text{where } x' = \frac{x - \bar{x}}{\text{std}(x)} \quad \text{and} \quad y' = \frac{y - \bar{y}}{\text{std}(y)}$$

$$\begin{aligned}x' &= \begin{matrix} -.866 & .866 & -.866 & .866 \end{matrix} \\y' &= \begin{matrix} .866 & -.866 & .866 & -.866 \end{matrix}\end{aligned}$$

$$\text{Corr}(x, y) = \frac{x' \cdot y'}{n-1} = \frac{-0.866 \cdot 0.866 + 0.866 \cdot -0.866 + -0.866 \cdot 0.866 + 0.866 \cdot -0.866}{3} = \frac{-3}{3} = -1$$

**Euclidean:**

$$\text{dist}(x, y) = \left( \sum_{k=1}^n |x_k - y_k|^r \right)^{\frac{1}{r}} \quad \text{where } n = 4, r = 2$$

$$\text{dist}(x, y) = \left( \sum_{k=1}^4 |x_k - y_k|^2 \right)^{\frac{1}{2}} = \left( |0-1|^2 + |1-0|^2 + |0-1|^2 + |1-0|^2 \right)^{\frac{1}{2}} = 2$$

**Jaccard:**

$$M_{01} = 2, \quad M_{10} = 2, \quad M_{11} = 0$$

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{0}{2 + 2 + 0} = 0$$

(c)

$$\begin{array}{r} x = 0 \quad -1 \quad 0 \quad 1 \\ y = 1 \quad 0 \quad -1 \quad 0 \end{array}$$

**Cosine:**

$$\cos(x, y) = (x \cdot y) / \|x\| \|y\| = 0 / 2 = 0$$

$$(x \cdot y) = 0 \cdot 1 + (-1) \cdot 0 + 0 \cdot (-1) + 1 \cdot 0 = 0$$

$$\|x\| = \sqrt{0^2 + (-1)^2 + 0^2 + 1^2} = \sqrt{2}, \quad \|y\| = \sqrt{1^2 + 0^2 + (-1)^2 + 0^2} = \sqrt{2}$$

**Correlation:**

$$\text{Corr}(x, y) = \frac{x' \cdot y'}{n-1} \quad \text{where } x' = \frac{x - \bar{x}}{\text{std}(x)} \quad \text{and} \quad y' = \frac{y - \bar{y}}{\text{std}(y)}$$

$$\begin{array}{r} x' = \quad 0 \quad -1.225 \quad 0 \quad 1.225 \\ y' = \quad 1.225 \quad 0 \quad -1.225 \quad 0 \end{array}$$

$$\text{Corr}(x, y) = \frac{x' \cdot y'}{n-1} = \frac{0 \cdot 1.225 + (-1.225) \cdot 0 + 0 \cdot (-1.225) + 1.225 \cdot 0}{3} = \frac{0}{3} = 0$$

**Euclidean:**

$$\text{dist}(x, y) = \left( \sum_{k=1}^n |x_k - y_k|^r \right)^{\frac{1}{r}} \quad \text{where } n = 4, r = 2$$

$$\text{dist}(x, y) = \left( \sum_{k=1}^4 |x_k - y_k|^2 \right)^{\frac{1}{2}} = \left( |0-1|^2 + |-1-0|^2 + |0+1|^2 + |1-0|^2 \right)^{\frac{1}{2}} = 2$$

(d)

$$\begin{array}{rcccccc} x & = & 1 & 1 & 0 & 1 & 0 & 1 \\ y & = & 1 & 1 & 1 & 0 & 0 & 1 \end{array}$$

**Cosine:**

$$\cos(x, y) = (x \cdot y) / \|x\| \|y\| = 3/4 = .75$$

$$(x \cdot y) = 1 \cdot 1 + 1 \cdot 1 + 0 \cdot 1 + 1 \cdot 0 + 0 \cdot 0 + 1 \cdot 1 = 3$$

$$\|x\| = \sqrt{1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = \sqrt{4} = 2,$$

$$\|y\| = \sqrt{1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2} = \sqrt{4} = 2$$

**Correlation:**

$$\text{Corr}(x, y) = \frac{x' \cdot y'}{n-1} \quad \text{where } x' = \frac{x - \bar{x}}{\text{std}(x)} \quad \text{and} \quad y' = \frac{y - \bar{y}}{\text{std}(y)}$$

$x'$	$y'$	$x'y'$
0.645497	0.645497	0.416667
0.645497	0.645497	0.416667
-1.29099	0.645497	-0.833333
0.645497	-1.29099	-0.833333
-1.29099	-1.29099	1.666667
0.645497	0.645497	0.416667
	sum(x'y')	1.25

$$\text{Corr}(x, y) = \frac{x' \cdot y'}{n-1} = \frac{1.25}{5} = .25$$

**Jaccard:**

$$M_{01} = 1, M_{10} = 1, M_{11} = 3$$

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{3}{1+1+3} = \frac{3}{5} = 0.6$$

(e)

$$\begin{array}{rcccccc} x & = & 2 & -1 & 0 & 2 & 0 & -3 \\ y & = & -1 & 1 & -1 & 0 & 0 & -1 \end{array}$$

**Cosine:**

$$\cos(x, y) = (x \cdot y) / \|x\| \|y\| = \frac{-6}{6\sqrt{2}} = \frac{-\sqrt{2}}{2} = -0.70711$$

$$(x \cdot y) = 2 \cdot -1 + -1 \cdot 1 + 0 \cdot -1 + 2 \cdot 0 + 0 \cdot 0 + -3 \cdot -1 = -6$$

$$\|x\| = \sqrt{2^2 + (-1)^2 + 0^2 + 2^2 + 0^2 + (-3)^2} = \sqrt{18} \approx 4.2426,$$

$$\|y\| = \sqrt{(-1)^2 + 1^2 + (-1)^2 + 0^2 + 0^2 + (-1)^2} = \sqrt{4} = 2$$

**Correlation:**

$$\text{Corr}(x, y) = \frac{x' \cdot y'}{n-1} \quad \text{where } x' = \frac{x - \bar{x}}{\text{std}(x)} \quad \text{and} \quad y' = \frac{y - \bar{y}}{\text{std}(y)}$$

$x'$	$y'$	$x'y'$
0.645497	-0.8165	-0.52705
-1.29099	1.632993	-2.10819
-0.6455	-0.8165	0.527046
0.645497	0.408248	0.263523
-0.6455	0.408248	-0.26352
1.290994	-0.8165	-1.05409
	sum(x'y')	-3.16228

$$\text{Corr}(x, y) = \frac{x' \cdot y'}{n-1} = \frac{-3.16228}{5} = -0.6325$$

7.

Instance	Type of Iris	Length	Binarization Method		
			(1)	(2)	
1	Setosa	4.5	short	short	Interval 1 = short Interval 2 = long  Class 1 = Setosa Class 2 = Virginica
2	Setosa	4.7	short	short	
3	Setosa	4.8	short	short	
4	Setosa	5.3	long	short	
5	Setosa	5.5	long	short	
6	Virginica	5.1	long	short	
7	Virginica	5.7	long	long	
8	Virginica	6.9	long	long	
9	Virginica	7.6	long	long	
10	Virginica	7.7	long	long	

$$e = \sum_{i=1}^n w_i e_i \quad \text{where} \quad w_i = \frac{m_i}{m} \quad e_i = -\sum_{j=1}^k p_{ij} \log_2 p_{ij} \quad p_{ij} = \frac{m_{ij}}{m_i}$$

### Method (1)

$$p_{11} = \frac{m_{11}}{m_1} = \frac{3}{3} = 1$$

$$e_1 = -(1 \log_2 1 + 0 \log_2 0) = 0$$

$$p_{12} = \frac{m_{12}}{m_1} = \frac{0}{3} = 0$$

$$p_{21} = \frac{m_{21}}{m_1} = \frac{2}{7} = 0.2857$$

$$e_2 = -\left(\frac{2}{7} \log_2 \frac{2}{7} + \frac{5}{7} \log_2 \frac{5}{7}\right) = -(-0.86312) = 0.86312$$

$$p_{22} = \frac{m_{22}}{m_1} = \frac{5}{7} = 0.7143$$

$$e = \sum_{i=1}^n w_i e_i = \frac{m_1}{m} e_1 + \frac{m_2}{m} e_2 = \frac{3}{10}(0) + \frac{7}{10}(0.86312) = 0.60418$$

## Method (2)

$$p_{11} = \frac{m_{11}}{m_1} = \frac{5}{6} = 0.83333$$

$$p_{12} = \frac{m_{11}}{m_1} = \frac{1}{6} = 0.16667$$

$$e_1 = -\left(\frac{5}{6}\log_2 \frac{5}{6} + \frac{1}{6}\log_2 \frac{1}{6}\right) = -(-0.65002) = 0.65002$$

$$p_{21} = \frac{m_{11}}{m_1} = \frac{0}{4} = 0$$

$$p_{22} = \frac{m_{11}}{m_1} = \frac{4}{4} = 1$$

$$e_2 = -(0\log_2 0 + 1\log_2 1) = 0$$

$$e = \sum_{i=1}^n w_i e_i = \frac{m_1}{m} e_1 + \frac{m_2}{m} e_2 = \frac{6}{10}(0.65002) + \frac{4}{10}(0) = 0.39001$$