Regression Tools

$\bar{y} = \mathbf{85.0}$

$\bar{x} = \mathbf{24.8}$

$SS(x) = \mathbf{1,743.64}$

$SS(y) = \mathbf{16,976.00}$

$SS(xy) = \mathbf{2,766.00}$

So we can calculate

$$Slope = b = \frac{SS(xy)}{SS(x)} = \frac{2,766.00}{1,743.64} = 1.59$$

$$Intercept = a = \bar{y} - b\bar{x}$$
$$= 85.0 + 1.59(24.8)$$
$$= 45.54$$

The straight line depicting the regression relationship of y on x is

$$\hat{y} = a + bx$$
$$\hat{y} = 45.54 + 1.59x$$

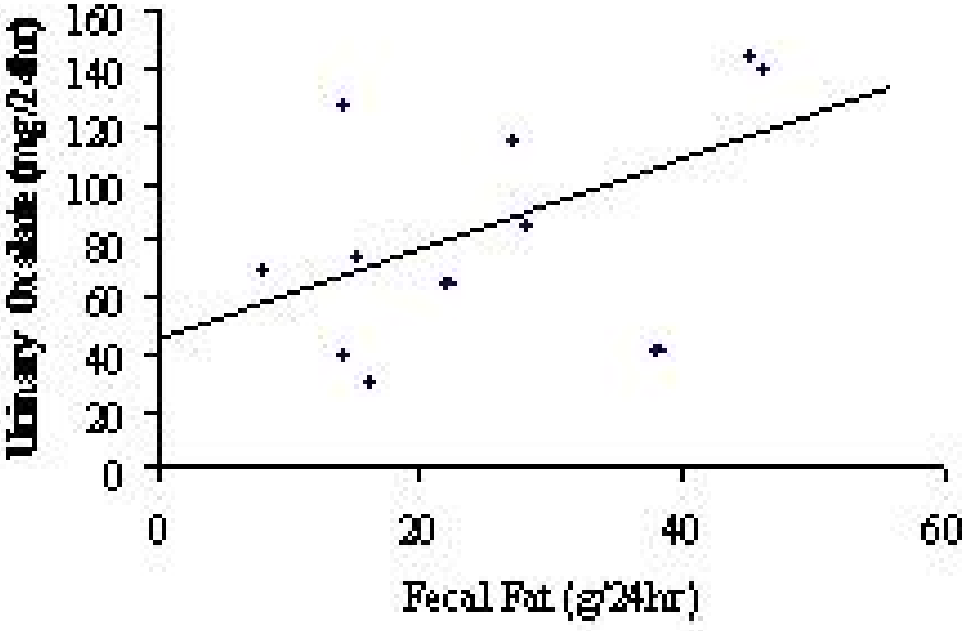
At $x = 40$, the regression estimate for y is:

$$\hat{y} = 45.54 + 1.59(40) = 109.14$$

With this information, we can add the regression line

$$\hat{y} = 45.54 + 1.59x$$

to the scatter plot, as shown below.



$$\hat{y} = 45.54 + 1.59(x = 45) = 117.09$$

Hypothesis test for regression of urinary oxalate on fecal fat

1. *The Hypothesis:* $H_0: \beta = 0$ vs $H_1: \beta \neq 0$

2. *The α level:* $\alpha = 0.05$

3. *The assumptions:* Random normal samples for y-variable from populations defined by x-variable

4. *The test statistic:* ANOVA as specified by

ANOVA				
Source	df	SS	MS	F
Regression	1	bSS(xy)	SS(Reg)/1	MS(Reg)/MS(Res)
Residual	n-2	SS(Res)[a]	SS(Res)/(n-2)	
Total	n-1	SS(y)		

[a] SS(Residual) = SS(y) − SS(Regression)

5. *The critical region:* Reject $H_0: \beta = 0$ if the value calculated for F is greater than $F_{0.95}(1,9) = 5.12$

6. *The result:*

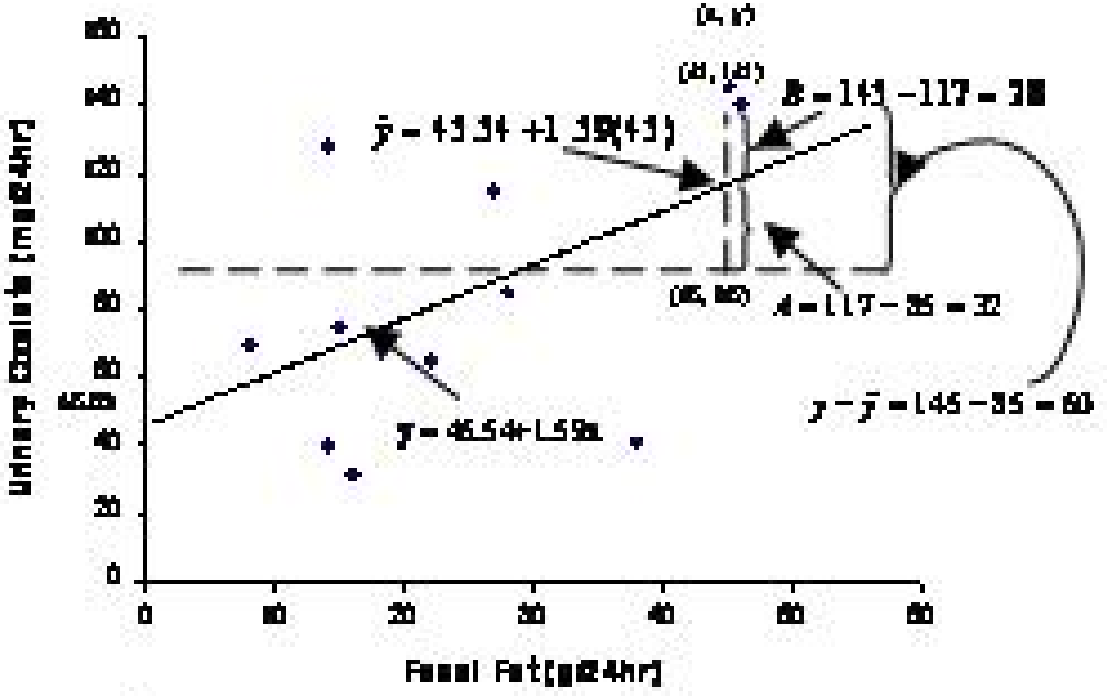
$SS(Reg) = bSS(xy) = 1.59(2,766.00) = 4,397.94$

$SS(Total) = SS(y) = 16,976.00$

$SS(Res) = 16,976.00 - 4,397.94 = 12,578.06$

ANOVA				
Source	df	SS	MS	F
Regression	1	4,397.94	4,397.94	3.15
Residual	9	12,578.06	1,397.56	
Total	10	16,976.00		

7. *The conclusion:* Accept $H_0: \beta = 0$ since $F < 5.12$



$R^2 = 0.2585$

$\hat{y}$ = regression estimate $= 45.54 + 1.59x = 117$

C = A + B = $y - \bar{y}$ Total deviation $\Rightarrow$ SS(y)

A = $\hat{y} - \bar{y}$ Explained by line $\Rightarrow$ SS(Reg)

B = $y - \hat{y}$ Left over $\Rightarrow$ SS(Residual)

Correlation Tools

$$SS(x) = 1{,}743.64$$
$$SS(y) = 16{,}976.00$$
$$SS(xy) = 2{,}766.00$$

The estimate of the correlation coefficient is:

$$r_{xy} = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}}$$

$$r_{xy} = \frac{2{,}766.00}{\sqrt{[1{,}743.64][16{,}976.00]}} = 0.5084$$

The Correlation Coefficient

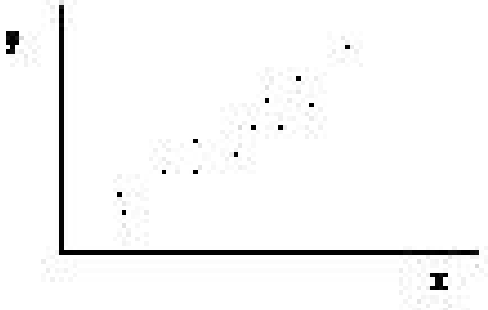
r measures linear association

r has values between $-1 \leq r \leq +1$

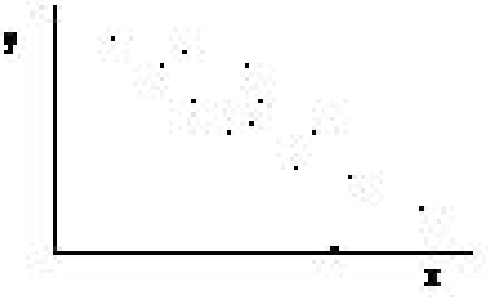
$r \approx +1$ implies strong positive linear association

$r \approx -1$ implies strong negative linear association

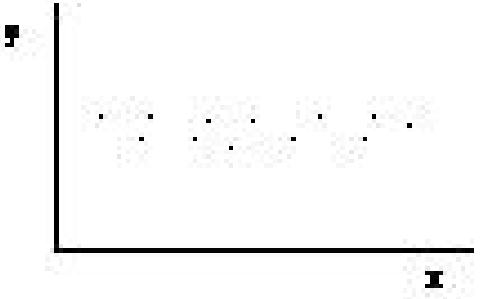
$r \approx 0$ implies no linear association



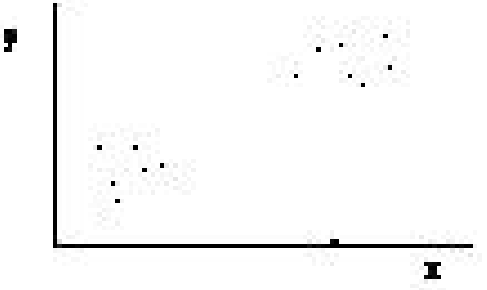
$r \approx +1$



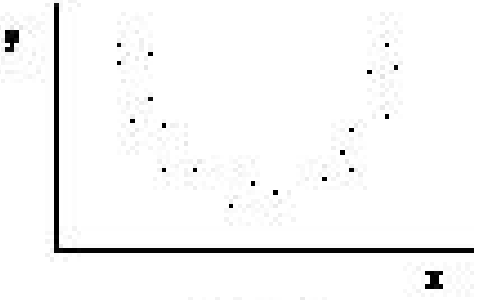
$r \approx -1$



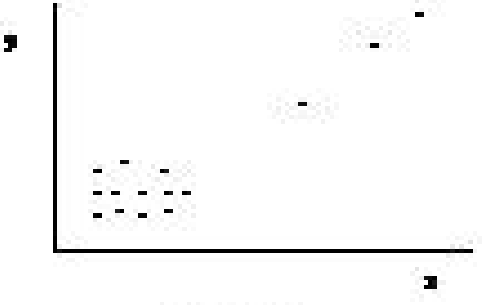
$r \approx 0$



$r \approx +1$



$r \approx 0$



$r > 0$

Correlation Hypothesis Testing

The hypothesis of interest deals with whether there is linear association between x and y. If there is no such association, we would have $\rho = 0$. Hence, the hypotheses of interest are:

$$H_0: \rho = 0 \quad \text{vs} \quad H_1: \rho \neq 0$$

which we can test by using the test statistic:

$$t = r\left[\frac{n-2}{1-r^2}\right]^{1/2} \approx t_{(n-2)}$$

Note that this calculation requires only the sample estimate r of the correlation coefficient ρ and the sample size n and that we need to use the t distribution with n - 2 degrees of freedom.

Test of correlation hypothesis for urinary oxalate and fecal fat, n = 11, r = 0.5084

1. *The Hypothesis:* $H_0: \rho = 0$ *vs* $H_1: \rho \neq 0$

2. *The α level:* $\alpha = 0.05$

3. *The assumptions:* Random sample from bivariate normal distribution

4. *The test statistic:*

$$t = r\left[\frac{n-2}{1-r^2}\right]^{1/2} \approx t_{(n-2)}$$

5. *The critical region:* Reject H_0: $\rho = 0$ if the value calculated for t is not between $\pm t_{0.975}(9) = 2.262$

6. *The result:* $r = 0.5084, \ n = 11$

$$t = 0.5084\left[\frac{9}{1-(0.5084)^2}\right]^{1/2}$$

$$= 0.5084\left[\frac{9}{1-0.2585}\right]^{1/2}$$

$$= 1.77$$

7. *The conclusion:* Accept H_0: $\rho = 0$ since $t = 1.77$ is between $\pm t_{0.975}(9) = 2.262$

Test of correlation hypothesis for Tono-Pen vs Goldman intraocular pressure, n = 40, r = 0.6574

1. *The hypothesis:* $H_0: \rho = 0$ **vs** $H_1: \rho \neq 0$

2. *The assumptions:* Random sample
bivariate normal distribution

3. *The* α*-level :* $\alpha = 0.05$

4. *The test statistic:* $t = r\sqrt{\frac{n-2}{1-r^2}}$

5. *The rejection region:* Reject H_0: $\rho = 0$, if t is not between $\pm t_{0.975}(38) = 2.02$

6. *The result:* $n = 40$, $r = 0.6574$, $r^2 = 0.44$,

$$t = 0.6574\sqrt{\frac{38}{1-0.6574^2}} = 0.66\sqrt{\frac{38}{0.56}} = 5.44$$

7. *The conclusion:* Reject H_0: $\rho = 0$

Since $t = 5.44$ is not between ± 2.02

Health Indicators and the Organization of Health Care Systems in Western Europe

ABSTRACT

Introduction

20 - 24

TABLE 3—Correlation Matrix of Selected Health and Socioeconomic Indicators of Western European Countries

	Health Care Expenditures per Capita	% Population Covered	Public Health Expenditures	Gini	Infant Mortality Rate	PYLL Males	PYLL Females	Life Expectancy Males	Life Expectancy Females
Gross domestic product per capita	.927	−.06	.02	−.33	−.647	−.28	−.10	−.49	.19
Health care expenditures per capita		−.13	−.11	−.42		−.16	−.07	−.58	.03
% population covered by health care system			.44	.01	.02	.00	.08	−.19	−.18
Public health analysis: costs as % of total health expenditures				−.15	−.13	−.60	.07	−.07	−.11
Gini coefficient					.241	.28	.09	−.22	−.14
Infant mortality rate						.37	.01	−.35	−.42
PYLL males							.01	−.89*	.08
PYLL females								−.01	−.95*
Life expectancy, males									.81*

Note: PYLL = potential years of life lost.
*Statistically significant at $p < .05$.

TABLE 5—Regression Models for Infant Mortality Rate as a Function of Gross Domestic Product, Health Care Expenditures, and Type of Health Care System, Western Europe

	p (SE)			
	Model I	Model II	Model III	Model IV
Gross domestic product per capita, US$	−0.294 (0.111)		−0.49 (0.111)	
Health care expenditures per capita, US$		−0.05 (0.016)		−0.26 (0.07)
Health care system: infant mortality (%) explained by type of health care system (R)			0.11 (0.06)	0.16 (0.07)
R^2	.09	.28	.01	.07
F^*	.01	.04	.01	.01

Note: All calculations exclude Luxembourg.

cluded by most cohort studies because of the high environmental risk-taking of these countries. In the case of the observed infant mortality rate, the predicted infant mortality rate on the basis of a given level of GDP. This ratio is expressed as a percentage and provides an estimate of infant mortality not explained by GDP in a given country.

The rate of observed to expected infant mortality rates according to GDP (Table 3, MR) is one of the infant mortality rate adjusted by the GDP. Its inverse produced the results are shown in Table 4 and in Figure 1.

Results

Mean values of all variables for the two types of health care systems are shown in Table 5. When we compared the means of socioeconomic indicators between the two groups, only two variables, GNP and health care expenditures, reached statistical significance.

Pearson correlations coefficients were statistically significant for health care expenditures and GNP, and for health variables and infant mortality (Table 3).

In simple regression models, health care expenditures and GDP were associated with life expectancy for females and inversely associated with potential years of life lost for females. However, they were not associated with the other health indicators for males. Health care expenditures explained 22% of the variability in potential years of life lost for females ($F = 7.93$, $P < .05$) and 37% of the expectancy for females ($F = 8.84$, $P < .05$). GDP explained 20% of the variability in potential life years lost for females ($F = 5.76$, $P < .05$) and 20% of life expectancy for females ($F = 4.96$, $P < .05$). Variables measuring income were explained significantly by these variables in these two health indicators.

In the analysis health care expenditures proved to be a better explanatory variable of infant mortality ($R^2 = .40$) than GDP ($R^2 = .24$). Type of health care system was included in models III and IV (Table 5), according to which infant mortality rates would be lower for countries having national health services compared with social security systems as a similar level of GDP (model III) and health care expenditures (model IV). The suggestion of this difference is consistent with the coefficient of 0.5, which is of approximately 0.1 (model III) or 0.2 (model IV).

Table 4 shows the best model that estimates the rate of observed to predicted infant mortality rates according to GDP. The introduction of health care expenditures (dummy) produces increases in health care expenditures would result in greater decreases in the infant mortality rate for countries with universal health services. The model predicts that of average health care expenditures ($1,000), the infant mortality rate would be about 10% lower in national health services. Figure 1 illustrates the resulting effect of the model on the ratio for the countries included in the study as health care expenditures increase, the decreasing variability of infant mortality

Test of correlation hypothesis for infant mortality rate and gross domestic product, n = 17, r = -0.64

example

1. *The hypothesis:* $H_0: \rho = 0$ vs $H_1: \rho \neq 0$

2. *The assumptions:* Random sample
bivariate normal distribution

3. *The α-level :* $\alpha = 0.05$

4. *The test statistic:* $t = r\sqrt{\frac{n-2}{1-r^2}}$

5. *The rejection region:* Reject H_0: $\rho = 0$, if t is not between $\pm t_{0.975}(15) = 2.13$

6. *The result:* $n = 17$, $r = -0.64$,

$$t = -0.64\sqrt{\frac{15}{1-(-0.64)^2}} = -0.64\sqrt{\frac{15}{0.59}} = -3.23$$

7. *The conclusion:* Reject H_0: $\rho = 0$

Since $t = -3.23$ is not between ± 2.1315

Test of correlation hypothesis for life expectancy for males and females, n = 17, r = 0.67

1. *The hypothesis:* $H_0: \rho = 0$ vs $H_1: \rho \neq 0$

2. *The assumptions:* Random sample
bivariate normal distribution

3. *The α-level :* $\alpha = 0.05$

4. *The test statistic:* $t = r\sqrt{\frac{n-2}{1-r^2}}$

5. *The rejection region:* Reject H_0: $\rho = 0$, if t is not between $\pm t_{0.025}(15) = 2.1315$

6. *The result:* $n = 17, r = 0.67,$

$$t = 0.67\sqrt{\frac{15}{1-0.67^2}} = 0.67\sqrt{\frac{15}{1-0.45}}$$

$$t = 0.67\sqrt{\frac{15}{0.55}} = 0.67\sqrt{27.27}$$

$$t = 0.67(5.22) \qquad t = 3.49$$

7. *The conclusion:* Reject H_0: $\rho = 0$
Since $t = 3.49$ is not between ± 2.1315

Social Capital, Income Inequality, and Mortality

ABSTRACT

Introduction

20 -30

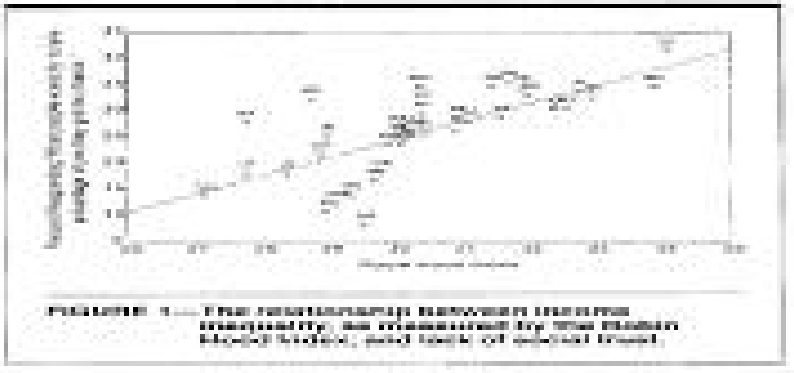


FIGURE 1—The relationship between income inequality, as measured by the Robin Hood Index, and lack of social trust.

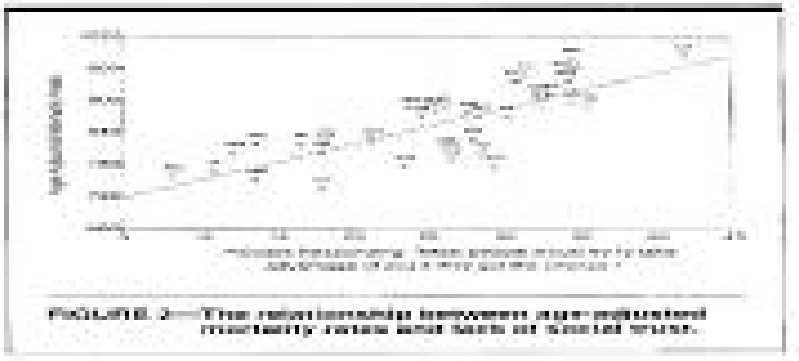


FIGURE 2—The relationship between age-adjusted mortality rates and lack of social trust.

poverty index in poorly wage income based and does not reflect other sources of income such as noncash benefits from Food Stamps, Medicaid, and public housing. Poverty thresholds are updated annually to reflect changes in the Consumer Price Index. The poverty variable used in the analyses represents the percentage of households in a given state below the federal poverty level (in 1990, an income of less than $13,359 for households with four family members[11]).

Measurement of State Mortality: An Overview

The age-adjusted mortality rates for each state in 1990 were obtained from the Compressed Mortality Files compiled by the National Center for Health Statistics of the Centers for Disease Control and Prevention (CDC). The data were obtained from the CDC's database via CDC WONDER software.[12]

All mortality rates were directly age standardized to the US population and expressed as the number of deaths per 100,000 persons (except in the case of infant mortality, for which deaths rates were expressed per 1000 live births). In addition to all-cause mortality, we examined the following major causes of death: coronary heart disease (International Classification of Diseases, 9th revision [ICD-9] codes 410 through 414); malignant neoplasms (ICD-9 codes 140 through 208); cerebrovascular disease (ICD-9 codes 430 through 438); and unintentional injuries (ICD-9 codes E800 through E949, V01 through V99).

TABLE 1—Correlations among Indicators of Social Capital, Poverty, Income Inequality, and Mortality

	Poverty	Mistrust	Group Membership	Group Membership-Work	Electoral Participation	Turnout
Mistrust	.63*					
Group membership	.73*	.69*				
Group membership-work	.69*	.69*	.88*			
Electoral participation	.49*	.55*	.71*	.74*		
Turnout	.52*	.70*	.71*	.69*	.79*	
Perceived lack of helpfulness	.65*	.81*	.71*	.64*	.79*	.91*

*Correlation is the percentage shown here. †Most analyses include only a random subsample of about 1 in every 10 observations.

Data Analysis

Ordinary least squares regression was used to examine the relationships of social capital indicators to mortality rates. Two sets of models were examined for each measure of interest. In the first set of models, we regressed the weighted social capital measures (e.g., weighted average group membership and weighted average social trust) against all-cause and cause-specific mortality rates. In the second set of models, we adjusted the regression models for state variations in prevalence of poverty. To examine the effects of inequality (as measured by the Robin Hood Index) and social capital on mortality, we carried out a path analysis[13] based on a causal model in which inequality affects mortality through its impact on social capital.

Results

Correlations among Social Capital Measures and Between Income Inequality and Social Capital

The four indicators of social capital—lack of participation in civic associations and the weighted proportions of respondents who agreed that "most people would try to take advantage of you if they got a chance," "most can't be bothered to help," or "people rarely look out for others"—were highly correlated with each other (Table 1). Since these variables may not represent an exhaustive list of "social capital" indica-

For Table 1, Correlation between Mortality and Social Mistrust, n = 39, r = 0.79

1. *The hypothesis:* $H_0: \rho = 0$ vs $H_1: \rho \neq 0$

2. *The assumptions:* Random sample
bivariate normal distribution

3. *The α level :* $\alpha = 0.05$

4. *The test statistic:* $t = r\sqrt{\frac{n-2}{1-r^2}}$

5. *The rejection region:* Reject H_0: $\rho = 0$, if t is not between $\pm t_{0.975}(37) = 2.02$

6. *The result:* $n = 39$, $r = 0.79$, $r^2 = 0.6241$,

$$t = 0.79\sqrt{\frac{37}{1-0.79^2}} = 0.79\sqrt{\frac{37}{0.3759}} = 7.8$$

7. *The conclusion:* Reject H_0: $\rho = 0$
Since $t = 7.8$ is not between ± 2.02

The Evolving Epidemiology of Chlamydial and Gonococcal Infections in Response to Control Programs in Winnipeg, Canada

ABSTRACT

20 - 34

Methods

Descriptive Epidemiology

Results

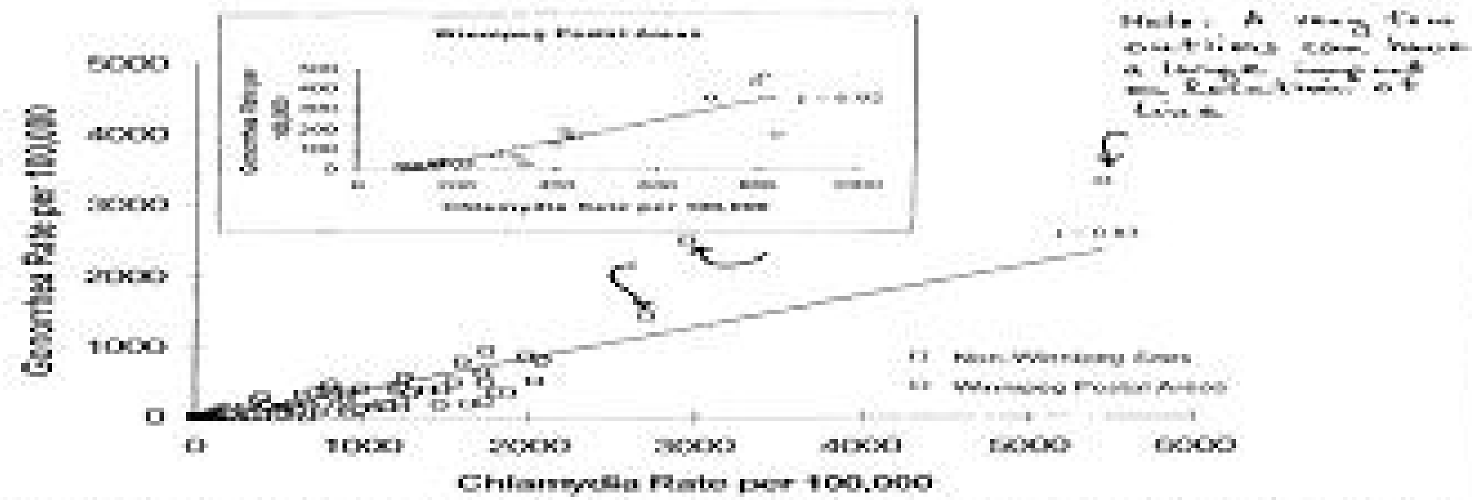


FIGURE 1—Scatterplot of average annual age- and sex-standardized incidence rates of chlamydia and gonorrhea, by Manitoba geographic site, 1991 through 1995.

20 - 35

Hypothesis test for correlation between gonorrhea rate and chlamydia rate

1. *The hypothesis:* $H_0: \rho = 0$ vs $H_1: \rho \neq 0$

2. *The assumptions:* random sample
bivariate normal distribution

3. *The α-level :* $\alpha = 0.05$

4. *The test statistic:* $$t = r\sqrt{\frac{n-2}{1-r^2}}$$

5. *The rejection region:* Reject H_0: $\rho = 0$, if t is not between $\pm t_{0.025}(320) \approx 2.00$

6. *The result:* $n = 332$, $r = 0.83$,

$$t = 0.83\sqrt{\frac{320}{1-0.83^2}} = 0.83\sqrt{\frac{320}{1-0.69}}$$

$$t = 0.83\sqrt{\frac{320}{0.31}} = 0.83\sqrt{1032.26}$$

$$t = 0.83(32.13) = 26.67$$

7. *The conclusion:* Reject H_0: $\rho = 0$

Since $t = 26.67$ is not between

Module 20: Correlation

This module focuses on the calculating, interpreting and testing hypotheses about the Pearson Product Moment Correlation Coefficient.

Correlation

In module 19, we examined how two variables, x and y, relate to each other by using the simple linear regression tool. In that context, x is the independent variable and y is the dependent variable. Typical examples for the independent variable include measures of time, including age; whereas, typical examples for the dependent variable are continuous measurements such as blood cholesterol level. The general assumption is that there are separate normal distributions of the dependent variable y for each value of the independent variable x. Further, we need to assume that these separate normal distributions for the dependent variable all have the same population variance.

Clearly these assumptions are quite restrictive in that we are often interested in the relationship between two variables, x and y, where it is not at all clear which should be labeled the independent variable and which the dependent one. An example is the relationship between blood cholesterol level and blood pressure level.

For this situation, we have another tool we can use to measure and test hypotheses about the relationship between these two variables. The tool is called correlation and we focus here only on what is usually called the Pearson Product Moment Correlation Coefficient. There are other measures of correlation which we will not discuss here. There are restrictions for the use of this correlation tool as well, which include the basic assumption that the x and y variables together have a joint frequency distribution which is called the bivariate normal distribution. This distribution looks like a three-dimensional bell in a manner similar to the way a normal distribution for one variable looks like a cross section of a bell.

The degree of association or correlation between two variables is measured by the *correlation coefficient*. This is done in a manner similar to that for other population parameters and estimates of these parameters obtained by calculating statistics from samples. That is, there is a value for the population parameter for this coefficient which is estimated by selecting a random sample and calculating the appropriate coefficient using the data from this sample. We can also use the information from the sample to test hypotheses about the population.

The population parameter for the Pearson Product Moment Correlation Coefficient is defined as

$$\rho_{xy} = \frac{\sum (x-\mu_x)(y-\mu_y)}{\sqrt{\sum (x-\mu_x)^2 \sum (y-\mu_y)^2}}$$

which is typically called rho, for the Greek letter it represents.

The estimate of ρ calculated from the sample data is the statistic

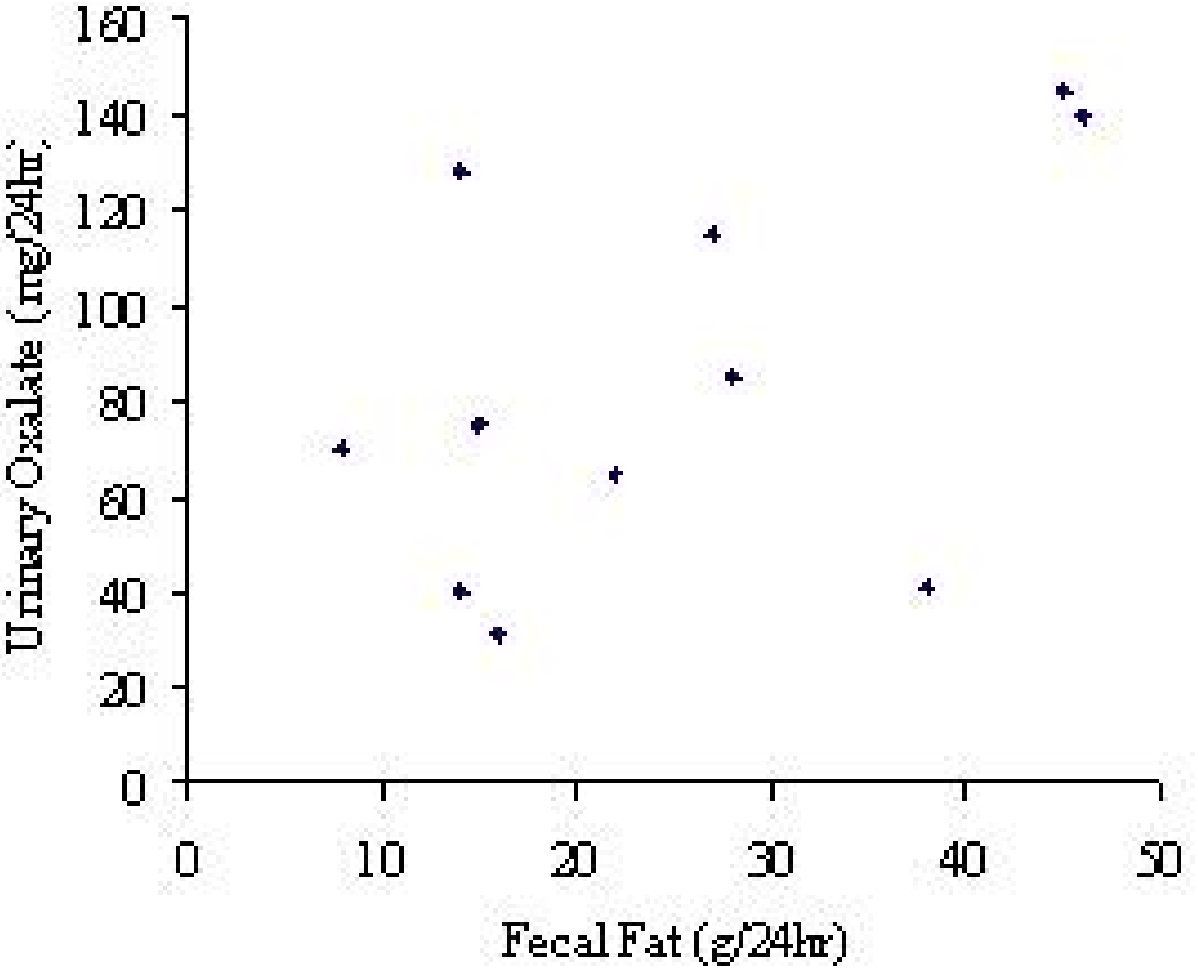
$$r_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

$$r_{xy} = \frac{\sum xy - (\sum x)(\sum y)/n}{\sqrt{[\sum x^2 - (\sum x)^2/n][\sum y^2 - (\sum y)^2/n]}}$$

$$= \frac{SS(xy)}{\sqrt{SS(x)SS(y)}}$$

Fecal Fat (g/24 hr) and Urinary Oxalate (mg/24 hr) secreted by a random sample of n = 11 persons

	Fecal Fat (g/24hr)	UrinaryOxalate (mg/24hr)
Patient	x	y
1	16	31
2	14	40
3	38	41
4	8	70
5	15	75
6	22	65
7	28	85
8	27	115
9	14	128
10	45	145
11	46	140
Sum	273	935
Mean	24.8	85.0


Urinary Oxalate (mg/24hr)
160
140
120
100
80
60
40
20
0
0
10
20
30
40
50
Fecal Fat (g/24hr)

	Fecal Fat		Urinary Oxalate		
Person	x	x^2	y	y^2	xy
1	16	256	31	961	496
2	14	196	40	1,600	560
3	38	1,444	41	1,681	1,558
4	8	64	70	4,900	560
5	15	225	75	5,625	1,125
6	22	484	65	4,225	1,430
7	28	784	85	7,225	2,380
8	27	729	115	13,225	3,105
9	14	196	128	16,384	1,792
10	45	2,025	145	21,025	6,525
11	46	2,116	140	19,600	6,440
Sum	273	8,519	935	96,451	25,971
Mean	24.8		85.0		
Sum^2/n	6,775.36		79,475.00		
SS	1,743.64		16,976.00		2,766.00
Variance	174.36		1,697.60		
SD	13.2		41.2		