Correlation

Major Points - Correlation

- Questions answered by correlation
- Scatterplots
- An example
- The correlation coefficient
- Other kinds of correlations
- Factors affecting correlations
- Testing for significance

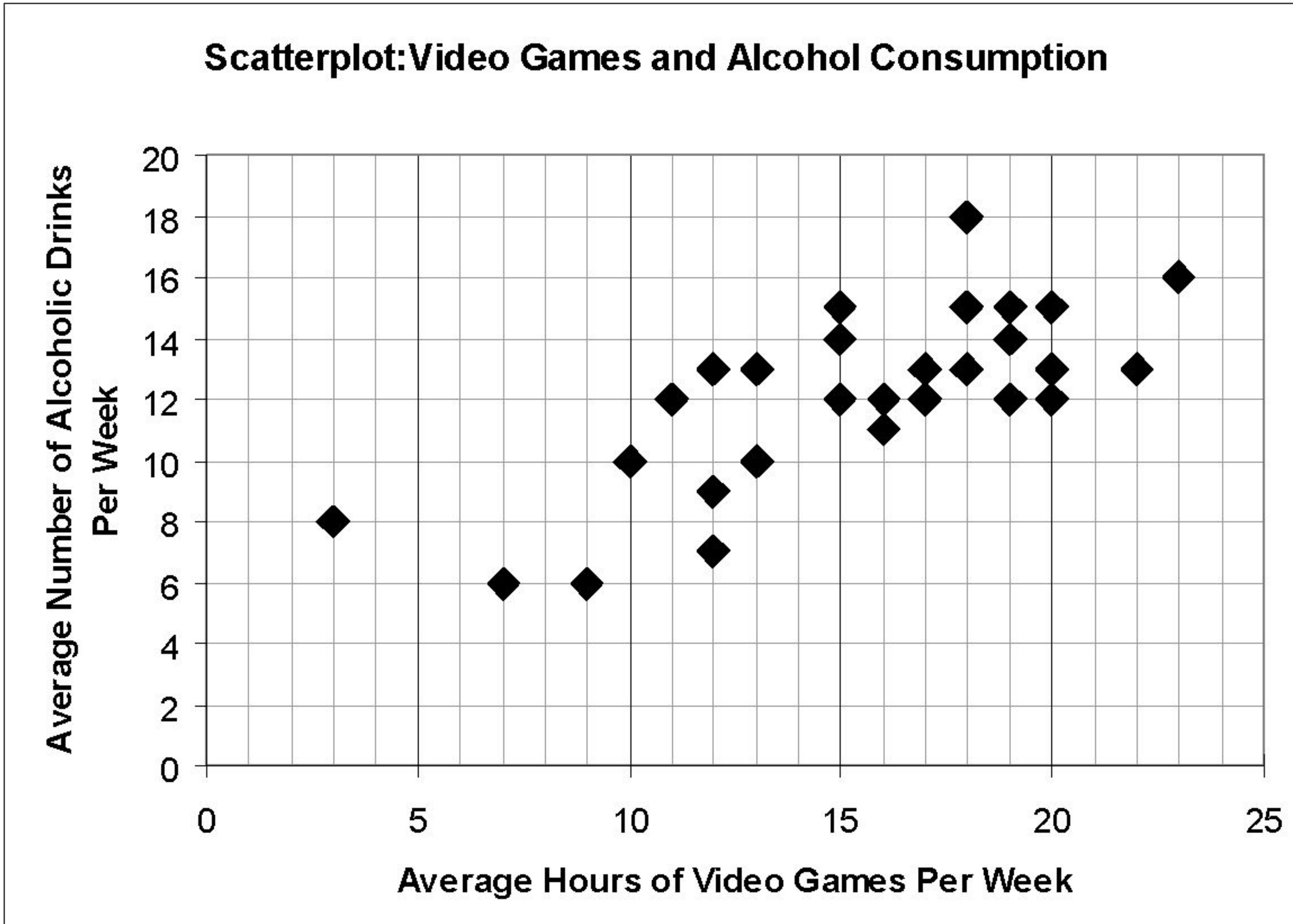
The Question

- Are two variables related?
  - Does one increase as the other increases?
    - e. g. skills and income
  - Does one decrease as the other increases?
    - e. g. health problems and nutrition
- How can we get a numerical measure of the degree of relationship?

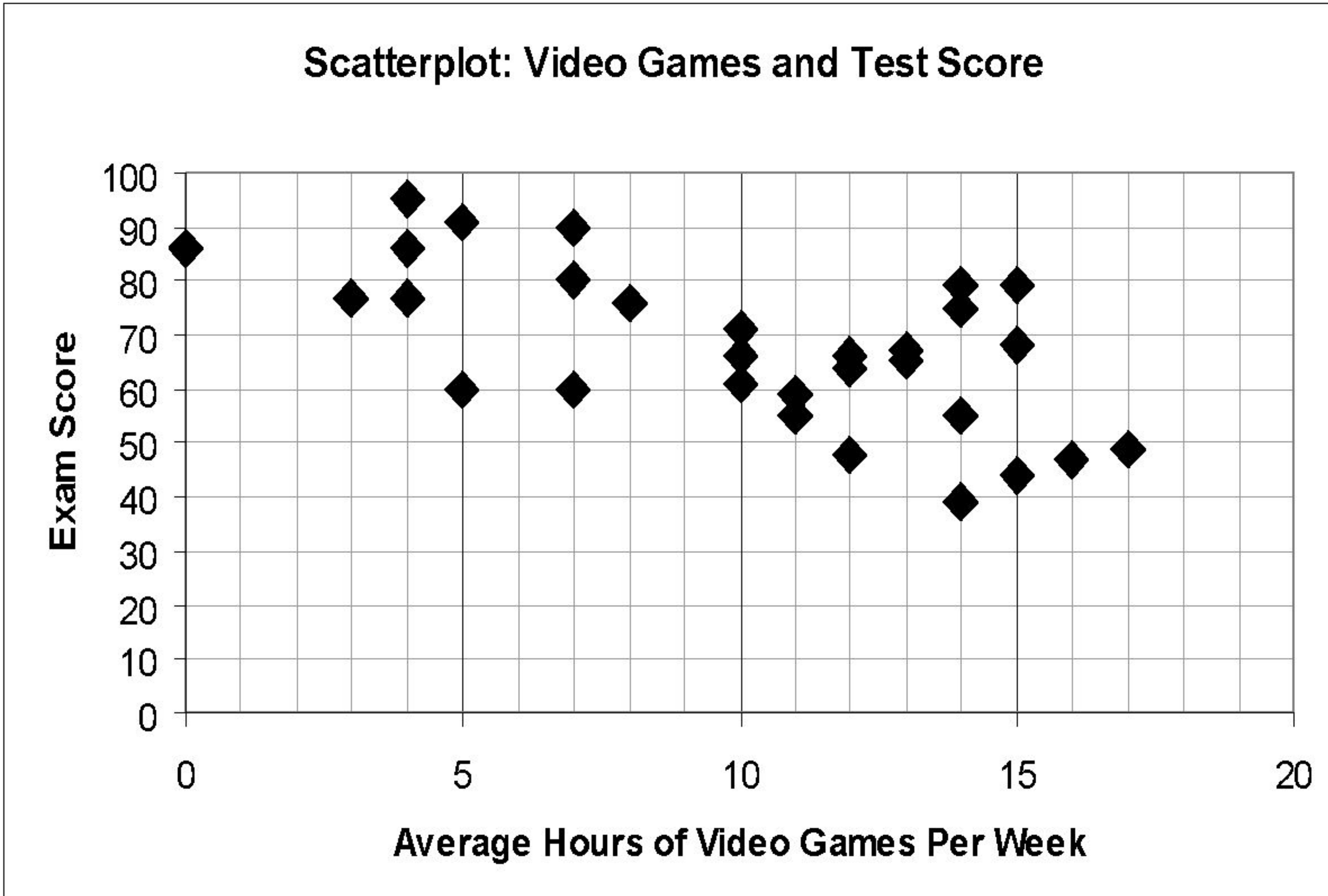
Scatterplots

- AKA scatter diagram or scattergram.
- Graphically depicts the relationship between two variables in two dimensional space.

Direct Relationship



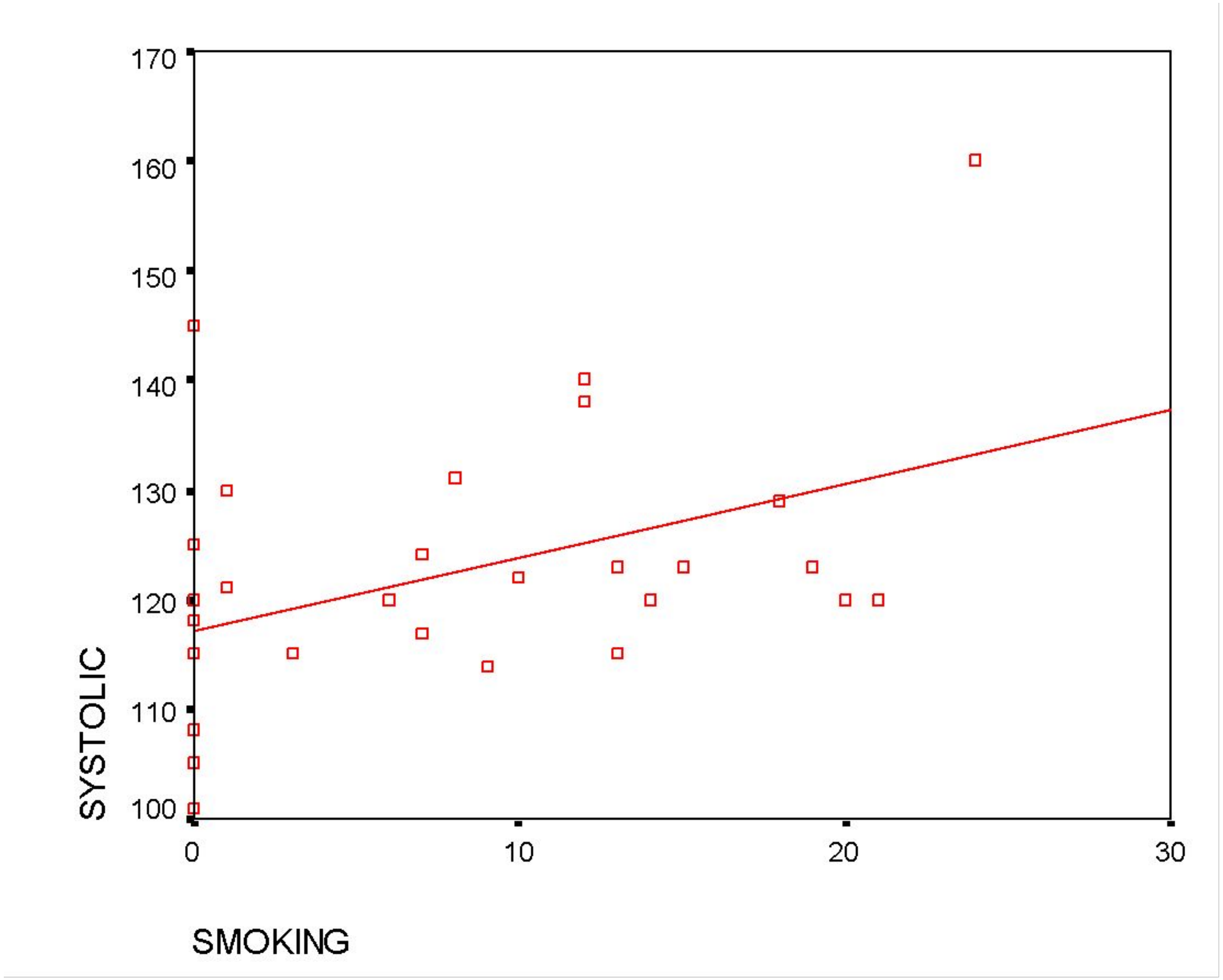
Inverse Relationship



An Example

- Does smoking cigarettes increase systolic blood pressure?
- Plotting number of cigarettes smoked per day against systolic blood pressure
  - Fairly moderate relationship
  - Relationship is positive

Trend?



Smoking and BP

- Note relationship is moderate, but real.
- Why do we care about relationship?
  - What would conclude if there were no relationship?
  - What if the relationship were near perfect?
  - What if the relationship were negative?

Heart Disease and Cigarettes

- Data on heart disease and cigarette smoking in 21 developed countries (Landwehr and Watkins, 1987)
- Data have been rounded for computational convenience.
  - The results were not affected.

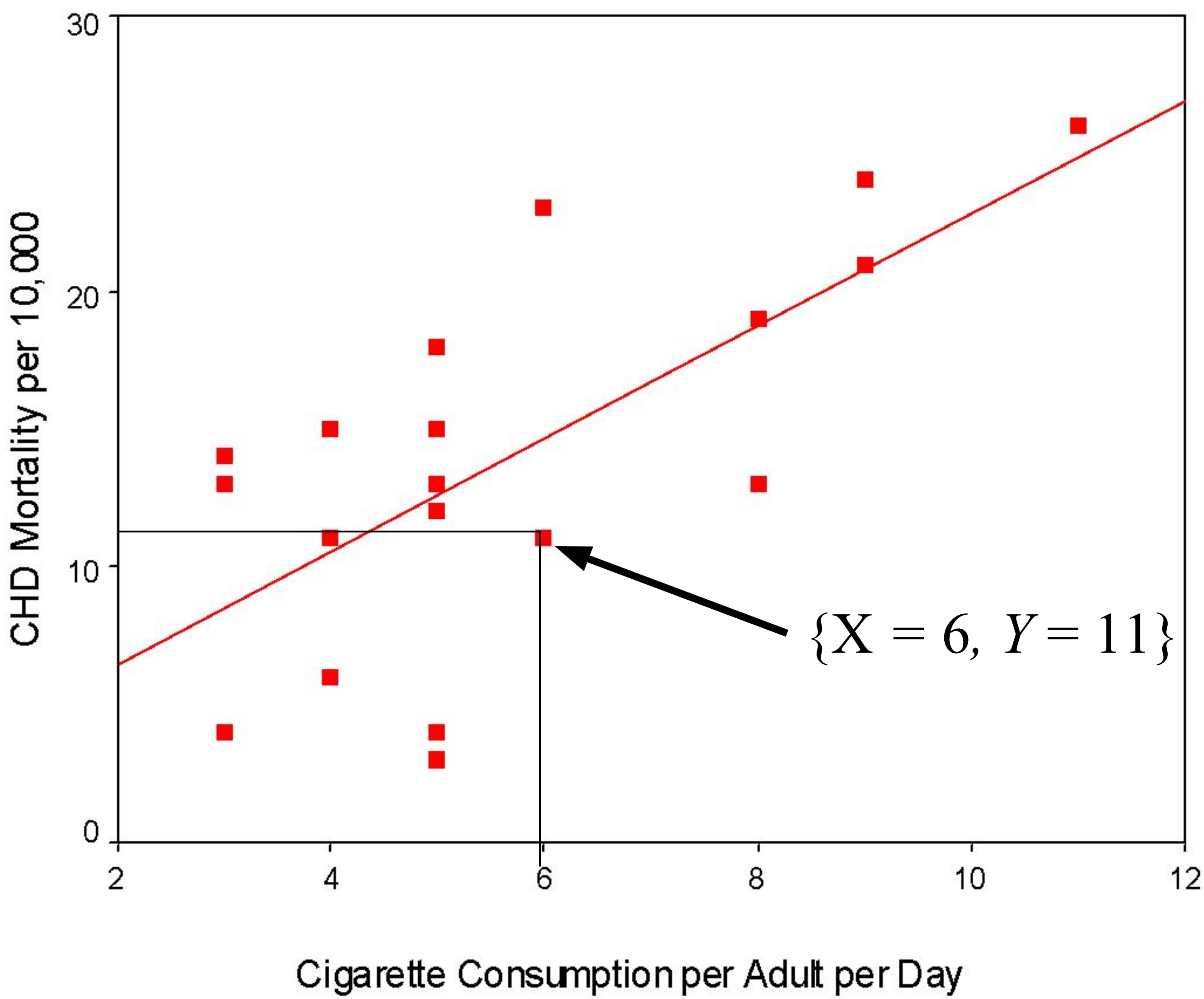
The Data

Surprisingly, the U.S. is the first country on the list--the country with the highest consumption and highest mortality.

Country	Cigarettes	CHD
1	11	26
2	9	21
3	9	24
4	9	21
5	8	19
6	8	13
7	8	19
8	6	11
9	6	23
10	5	15
11	5	13
12	5	4
13	5	18
14	5	12
15	5	3
16	4	11
17	4	15
18	4	6
19	3	13
20	3	4
21	3	14

Scatterplot of Heart Disease

- CHD Mortality goes on ordinate (Y axis)
  - Why?
- Cigarette consumption on abscissa (X axis)
  - Why?
- What does each dot represent?
- Best fitting line included for clarity



30
20
10
0
CHD Mortality per 10,000
{X = 6, Y = 11}
2
4
6
8
10
12
Cigarette Consumption per Adult per Day

What Does the Scatterplot Show?

- As smoking increases, so does coronary heart disease mortality.
- Relationship looks strong
- Not all data points on line.
  - This gives us “residuals” or “errors of prediction”
    - To be discussed later

Correlation

- Co-relation
- The relationship between two variables
- Measured with a correlation coefficient
- Most popularly seen correlation coefficient: Pearson Product-Moment Correlation

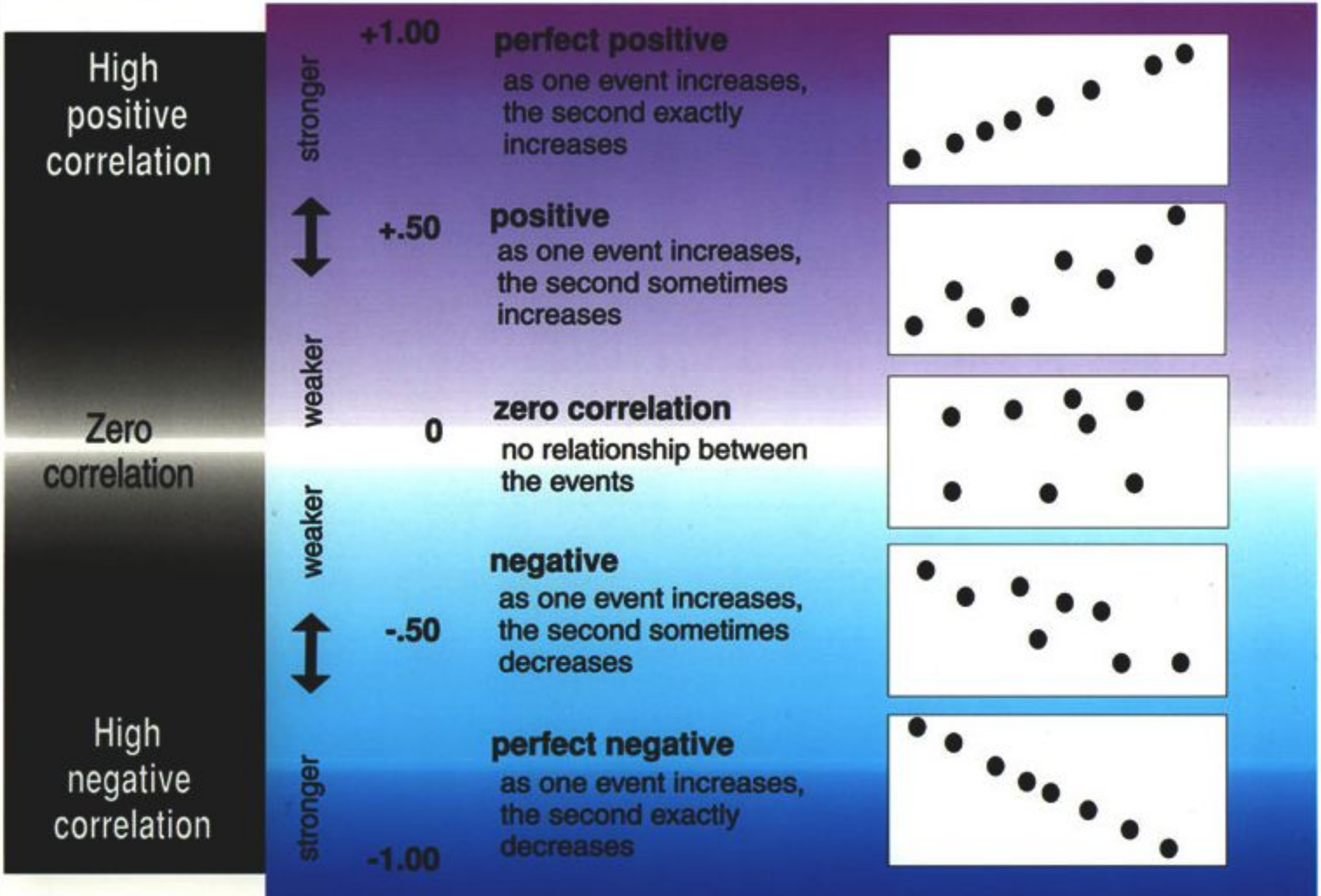
Types of Correlation

- Positive correlation
  - High values of X tend to be associated with high values of Y.
  - As X increases, Y increases
- Negative correlation
  - High values of X tend to be associated with low values of Y.
  - As X increases, Y decreases
- No correlation
- No consistent tendency for values on Y to increase or decrease as X increases

Correlation Coefficient

- A measure of degree of relationship.
- Between 1 and -1
- Sign refers to direction.
- Based on covariance
  - Measure of degree to which large scores on X go with large scores on Y, and small scores on X go with small scores on Y
  - Think of it as variance, but with 2 variables instead of 1 (What does that mean??)

Correlation



Covariance

- Remember that variance is:

$$Var_X = \frac{\Sigma(X - \overline{X})^2}{N-1} = \frac{\Sigma(X - \overline{X})(X - \overline{X})}{N-1}$$

- The formula for co-variance is:

$$Cov_{XY} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{N-1}$$

- How this works, and why?
- When would cov_{XY} be large and positive? Large and negative?

Example

Country	X (Cig.)	Y (CHD)	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})*(Y-\bar{Y})$
1	11	26	5.05	11.48	57.97
2	9	21	3.05	6.48	19.76
3	9	24	3.05	9.48	28.91
4	9	21	3.05	6.48	19.76
5	8	19	2.05	4.48	9.18
6	8	13	2.05	-1.52	-3.12
7	8	19	2.05	4.48	9.18
8	6	11	0.05	-3.52	-0.18
9	6	23	0.05	8.48	0.42
10	5	15	-0.95	0.48	-0.46
11	5	13	-0.95	-1.52	1.44
12	5	4	-0.95	-10.52	9.99
13	5	18	-0.95	3.48	-3.31
14	5	12	-0.95	-2.52	2.39
15	5	3	-0.95	-11.52	10.94
16	4	11	-1.95	-3.52	6.86
17	4	15	-1.95	0.48	-0.94
18	4	6	-1.95	-8.52	16.61
19	3	13	-2.95	-1.52	4.48
20	3	4	-2.95	-10.52	31.03
21	3	14	-2.95	-0.52	1.53
Mean	5.95	14.52			
SD	2.33	6.69			
Sum					222.44

Example

$$Cov_{cig.\&CHD} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{N-1} = \frac{222.44}{21-1} = 11.12$$

- What the heck is a covariance?
- I thought we were talking about correlation?

Correlation Coefficient

- Pearson's Product Moment Correlation
- Symbolized by *r*
- Covariance ÷ (product of the 2 SDs)

$$r = \frac{Cov_{XY}}{s_X s_Y}$$

- Correlation is a standardized covariance

Calculation for Example

- $\text{Cov}_{XY} = 11.12$
- $s_X = 2.33$
- $s_Y = 6.69$

$$r = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{11.12}{(2.33)(6.69)} = \frac{11.12}{15.59} = .713$$

Example

- Correlation = .713
- Sign is positive
  - Why?
- If sign were negative
  - What would it mean?
  - Would not alter the *degree* of relationship.

Other calculations

- Z-score method

$$r = \frac{\sum z_x z_y}{N-1}$$

- Computational (Raw Score) Method

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum Y^2 - (\sum Y)^2\right]}}$$

Other Kinds of Correlation

- Spearman Rank-Order Correlation Coefficient (r_{sp})
  - used with 2 ranked/ordinal variables
  - uses the same Pearson formula

Attractiveness	Symmetry
3	2
4	6
1	1
2	3
5	4
6	5

$r_{sp} = 0.77$

Other Kinds of Correlation

- Point biserial correlation coefficient (r_{pb})
  - used with one continuous scale and one nominal or ordinal or dichotomous scale.
  - uses the same Pearson formula

Attractiveness	Date?
3	0
4	0
1	1
2	1
5	1
6	0

$r_{pb} = -0.49$

Other Kinds of Correlation

- Phi coefficient (Φ)
  - used with two dichotomous scales.
  - uses the same Pearson formula

Attractiveness	Date?
0	0
1	0
1	1
1	1
0	0
1	1

$\Phi = 0.71$

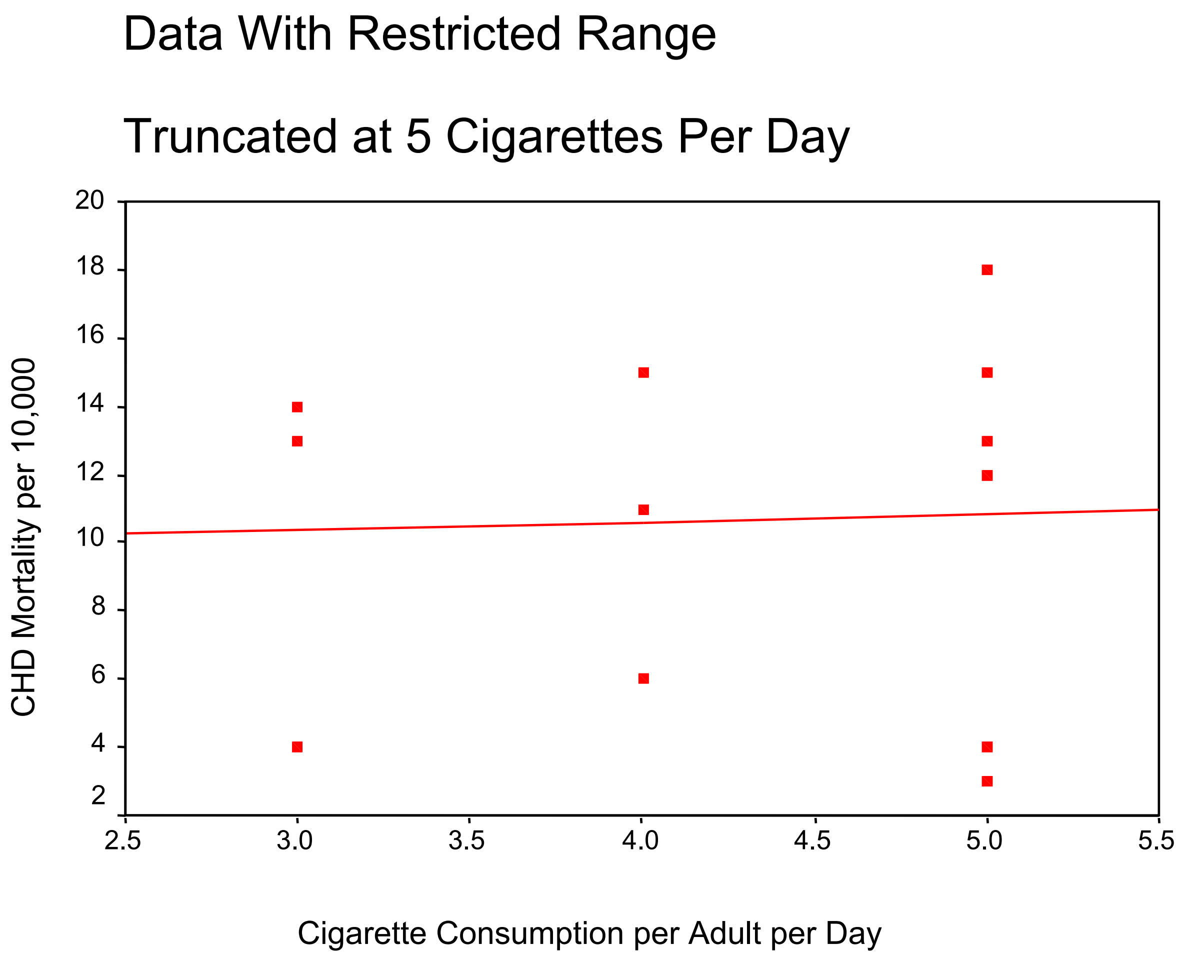
Factors Affecting *r*

- Range restrictions
  - Looking at only a small portion of the total scatter plot (looking at a smaller portion of the scores' variability) **decreases** *r*.
  - Reducing variability reduces *r*
- Nonlinearity
  - The Pearson r (and its relatives) measure the degree of **linear** relationship between two variables
  - If a strong non-linear relationship exists, r will provide a low, or at least inaccurate measure of the true relationship.

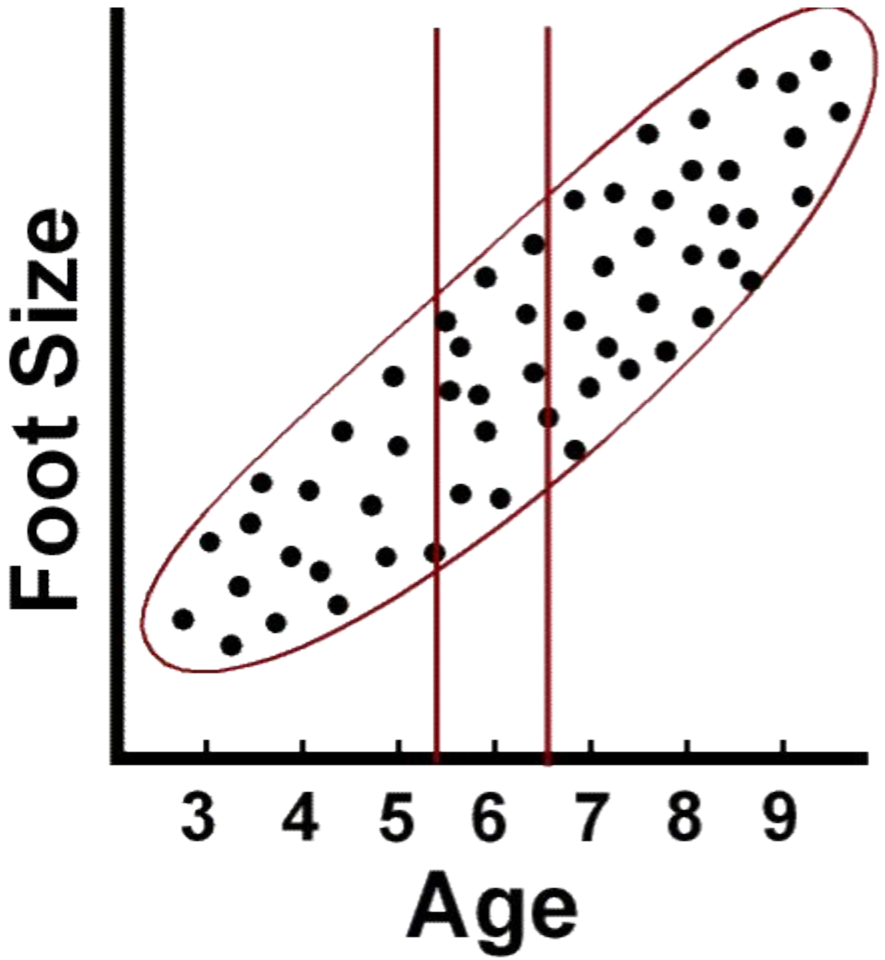
Factors Affecting *r*

- Heterogeneous subsamples
  - Everyday examples (e.g. height and weight using both men and women)
- Outliers
  - Overestimate Correlation
  - Underestimate Correlation

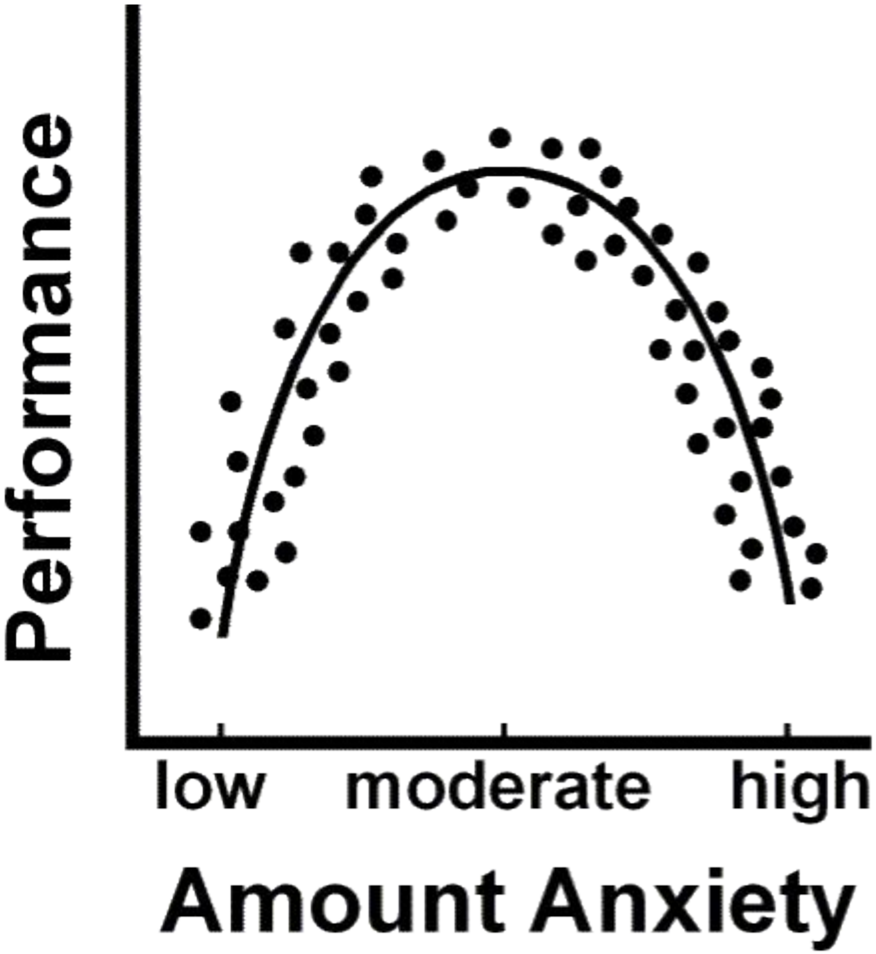
Countries With Low Consumptions



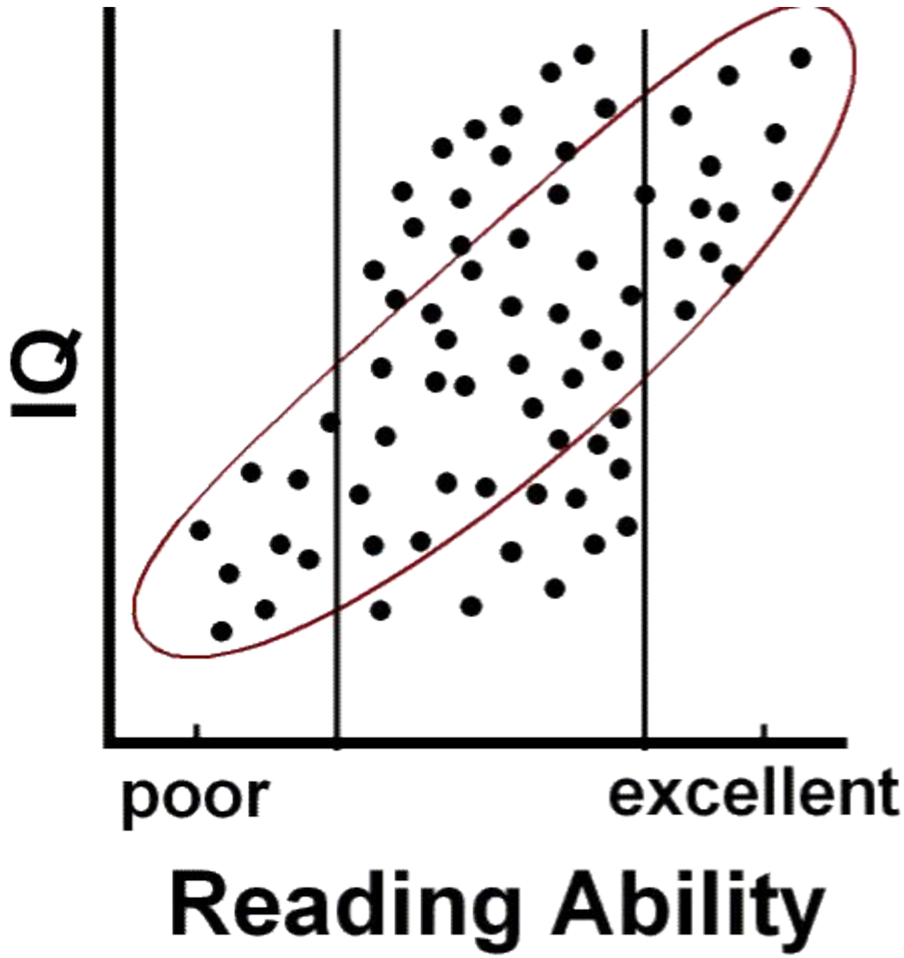
Truncation



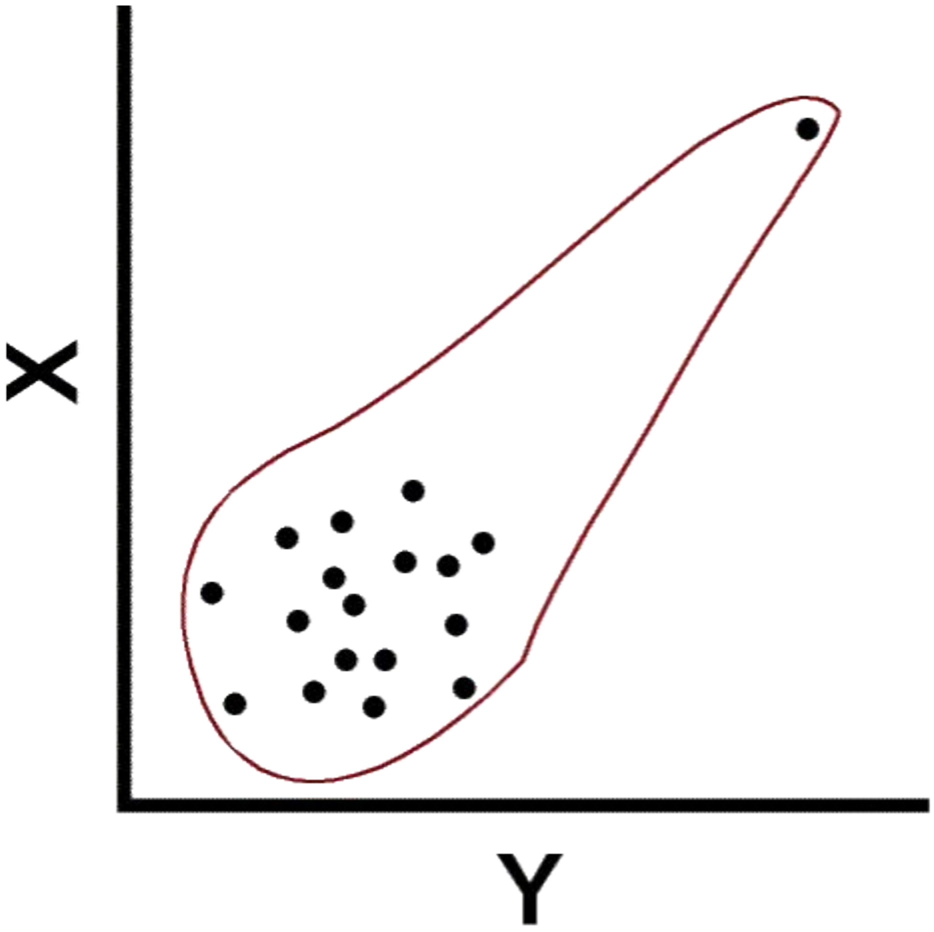
Non-linearity



Heterogenous samples



Outliers



Testing Correlations

- So you have a correlation. Now what?
- In terms of magnitude, how big is big?
  - Small correlations in large samples are "big."
  - Large correlations in small samples aren't always "big."
- Depends upon the magnitude of the correlation coefficient

AND

- The size of your sample.

Testing *r*

- Population parameter = ρ
- Null hypothesis H_0: ρ = 0
  - Test of linear independence
  - What would a true null mean here?
  - What would a false null mean here?
- Alternative hypothesis (H_1) $\rho \neq 0$
  - Two-tailed

Tables of Significance

- We can convert r to t and test for significance:

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

- Where DF = N-2

Tables of Significance

- In our example *r* was .71
- N-2 = 21 – 2 = 19

$$t = r\sqrt{\frac{N-2}{1-r^2}} = .71*\sqrt{\frac{19}{1-.71^2}} = .71*\sqrt{\frac{19}{.4959}} = 6.90$$

- T-crit (19) = 2.09
- Since 6.90 is larger than 2.09 reject $\rho = 0$.

Computer Printout

- Printout gives test of significance.

Correlations

		CIGARET	CHD
CIGARET	Pearson Correlation	1	.713**
	Sig. (2-tailed)	.	.000
	N	21	21
CHD	Pearson Correlation	.713**	1
	Sig. (2-tailed)	.000	.
	N	21	21

**. Correlation is significant at the 0.01 level (2-tailed).

Country	X (Cig.)	Y (CHD)	Y'	(Y - Y')	$(Y - Y')^2$	(Y' - Ybar)	(Y - Ybar)
1	11	26	24.829	1.171	1.371	106.193	131.699
2	9	21	20.745	0.255	0.065	38.701	41.939
3	9	24	20.745	3.255	10.595	38.701	89.795
4	9	21	20.745	0.255	0.065	38.701	41.939
5	8	19	18.703	0.297	0.088	17.464	20.035
6	8	13	18.703	-5.703	32.524	17.464	2.323
7	8	19	18.703	0.297	0.088	17.464	20.035
8	6	11	14.619	-3.619	13.097	0.009	12.419
9	6	23	14.619	8.381	70.241	0.009	71.843
10	5	15	12.577	2.423	5.871	3.791	0.227
11	5	13	12.577	0.423	0.179	3.791	2.323
12	5	4	12.577	-8.577	73.565	3.791	110.755
13	5	18	12.577	5.423	29.409	3.791	12.083
14	5	12	12.577	-0.577	0.333	3.791	6.371
15	5	3	12.577	-9.577	91.719	3.791	132.803
16	4	11	10.535	0.465	0.216	15.912	12.419
17	4	15	10.535	4.465	19.936	15.912	0.227
18	4	6	10.535	-4.535	20.566	15.912	72.659
19	3	13	8.493	4.507	20.313	36.373	2.323
20	3	4	8.493	-4.493	20.187	36.373	110.755
21	3	14	8.493	5.507	30.327	36.373	0.275
Mean	5.952	14.524					
SD	2.334	6.690					
Sum				0.04	440.757	454.307	895.247

$Y' = (2.04*X) + 2.37$

Example

Example

$$SS_{Total} = \sum (Y - \bar{Y})^2 = 895.247;\ df_{total} = 21 - 1 = 20$$

$$SS_{regression} = \sum (\hat{Y} - \bar{Y})^2 = 454.307;\ df_{regression} = 1 \text{ (only 1 predictor)}$$

$$SS_{residual} = \sum (Y - \hat{Y})^2 = 440.757;\ df_{residual} = 20 - 1 = 19$$

$$s^2_{total} = \frac{\sum (Y - \bar{Y})^2}{N - 1} = \frac{895.247}{20} = 44.762$$

$$s^2_{regression} = \frac{\sum (\hat{Y} - \bar{Y})^2}{1} = \frac{454.307}{1} = 454.307$$

$$s^2_{residual} = \frac{\sum (Y - \hat{Y})^2}{N - 2} = \frac{440.757}{19} = 23.198$$

$$Note: \sqrt{s^2_{residual}} = s_{Y - \hat{Y}}$$

Coefficient of Determination

- It is a measure of the percent of predictable variability

$$r^2 = \text{ the correlation squared}$$

or

$$r^2 = \frac{SS_{regression}}{SS_Y}$$

- The percentage of the total variability in Y explained by X

r^2 for our example

- $r = .713$
- $r^2 = .713^2 = .508$
- or $$r^2 = \frac{SS_{regression}}{SS_Y} = \frac{454.307}{895.247} = .507$$
- Approximately 50% in variability of incidence of CHD mortality is associated with variability in smoking.

Coefficient of Alienation

- It is defined as 1 - r^2 or

$$1 - r^2 = \frac{SS_{residual}}{SS_Y}$$

- Example

1 - .508 = .492

$$1 - r^2 = \frac{SS_{residual}}{SS_Y} = \frac{440.757}{895.247} = .492$$

r^2, SS and $s_{Y-Y'}$

- $r^2 * SS_{total} = SS_{regression}$
- $(1 - r^2) * SS_{total} = SS_{residual}$
- We can also use r^2 to calculate the standard error of estimate as:

$$s_{Y-\hat{Y}} = s_y \sqrt{(1-r^2)\left(\frac{N-1}{N-2}\right)} = 6.690 * \sqrt{(.492)\left(\frac{20}{19}\right)} = 4.816$$

Testing Overall Model

- We can test for the overall prediction of the model by forming the ratio:

$$\frac{s^2_{regression}}{s^2_{residual}} = F \text{ statistic}$$

- If the calculated F value is larger than a tabled value (F-Table) we have a significant prediction

Testing Overall Model

- Example $$\frac{s^2_{regression}}{s^2_{residual}} = \frac{454.307}{23.198} = 19.594$$
- F-Table – F critical is found using 2 things $df_{regression}$ (numerator) and $df_{residual.}$(demoninator)
- F-Table our F_{crit} (1,19) = 4.38
- 19.594 > 4.38, significant overall
- Should all sound familiar…

SPSS output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.713a	.508	.482	4.81640

a. Predictors: (Constant), CIGARETT

ANOVAb

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	454.482	1	454.482	19.592	.000a
	Residual	440.757	19	23.198		
	Total	895.238	20			

a. Predictors: (Constant), CIGARETT

b. Dependent Variable: CHD

Testing Slope and Intercept

- The regression coefficients can be tested for significance
- Each coefficient divided by it's standard error equals a t value that can also be looked up in a t-table
- Each coefficient is tested against 0

Testing the Slope

- With only 1 predictor, the standard error for the slope is:

$$se_b = \frac{s_{Y-\hat{Y}}}{s_X\sqrt{N-1}}$$

- For our Example:

$$se_b = \frac{4.816}{2.334\sqrt{21-1}} = \frac{4.816}{10.438} = .461$$

Testing Slope and Intercept

- These are given in computer printout as a *t* test.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.367	2.941		.805	.431
	Cigarette Consumption per Adult per Day	2.042	.461	.713	4.426	.000

a. Dependent Variable: CHD Mortality per 10,000

Testing

- The *t* values in the second from right column are tests on slope and intercept.
- The associated *p* values are next to them.
- The slope is significantly different from zero, but not the intercept.
- Why do we care?

Testing

- What does it mean if slope is not significant?
  - How does that relate to test on *r*?
- What if the intercept is not significant?
- Does significant slope mean we predict quite well?