Correlation

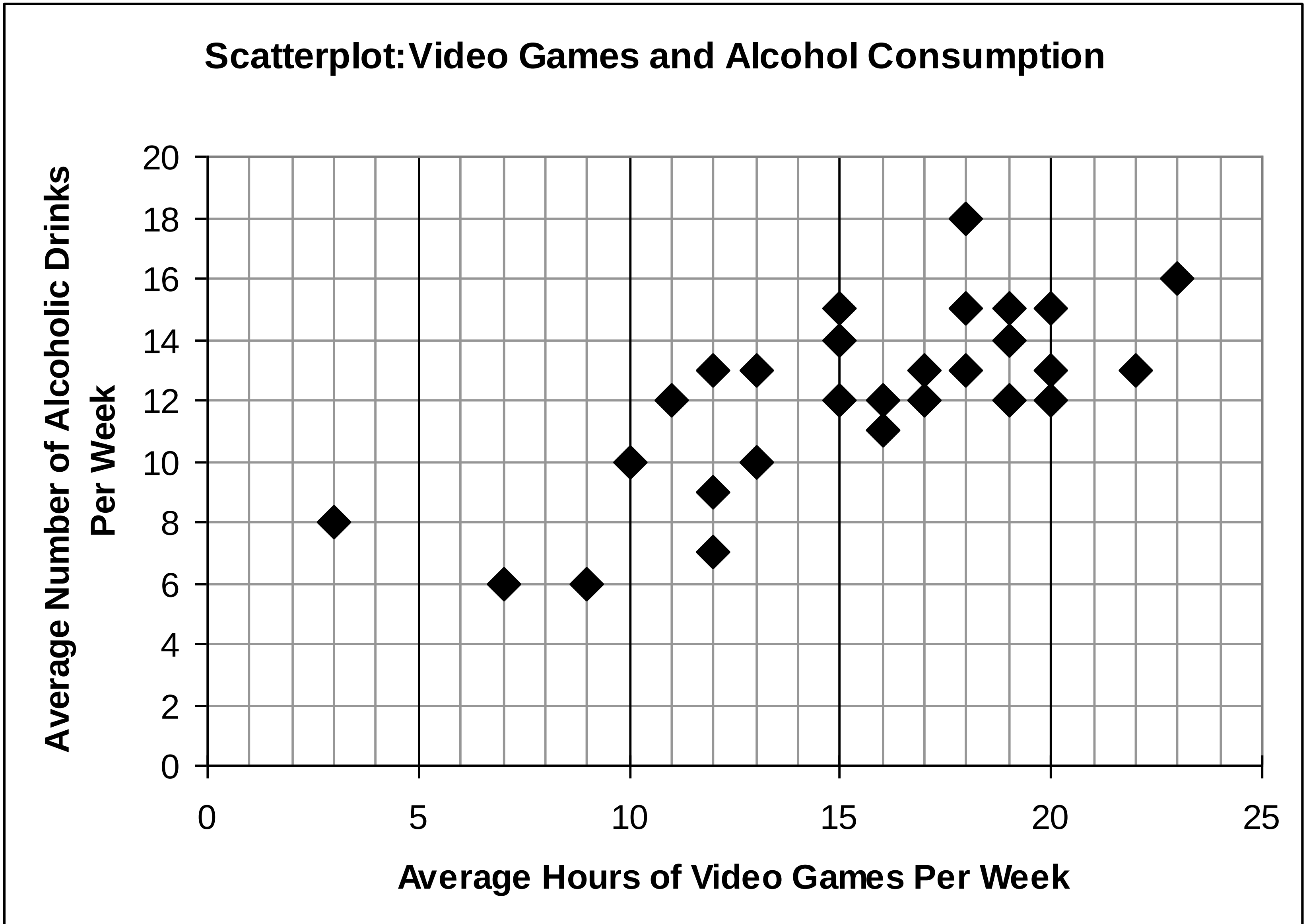
- Two variables are said to be correlated when change in the value of one variable results in the change in the value of other variable

- 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?

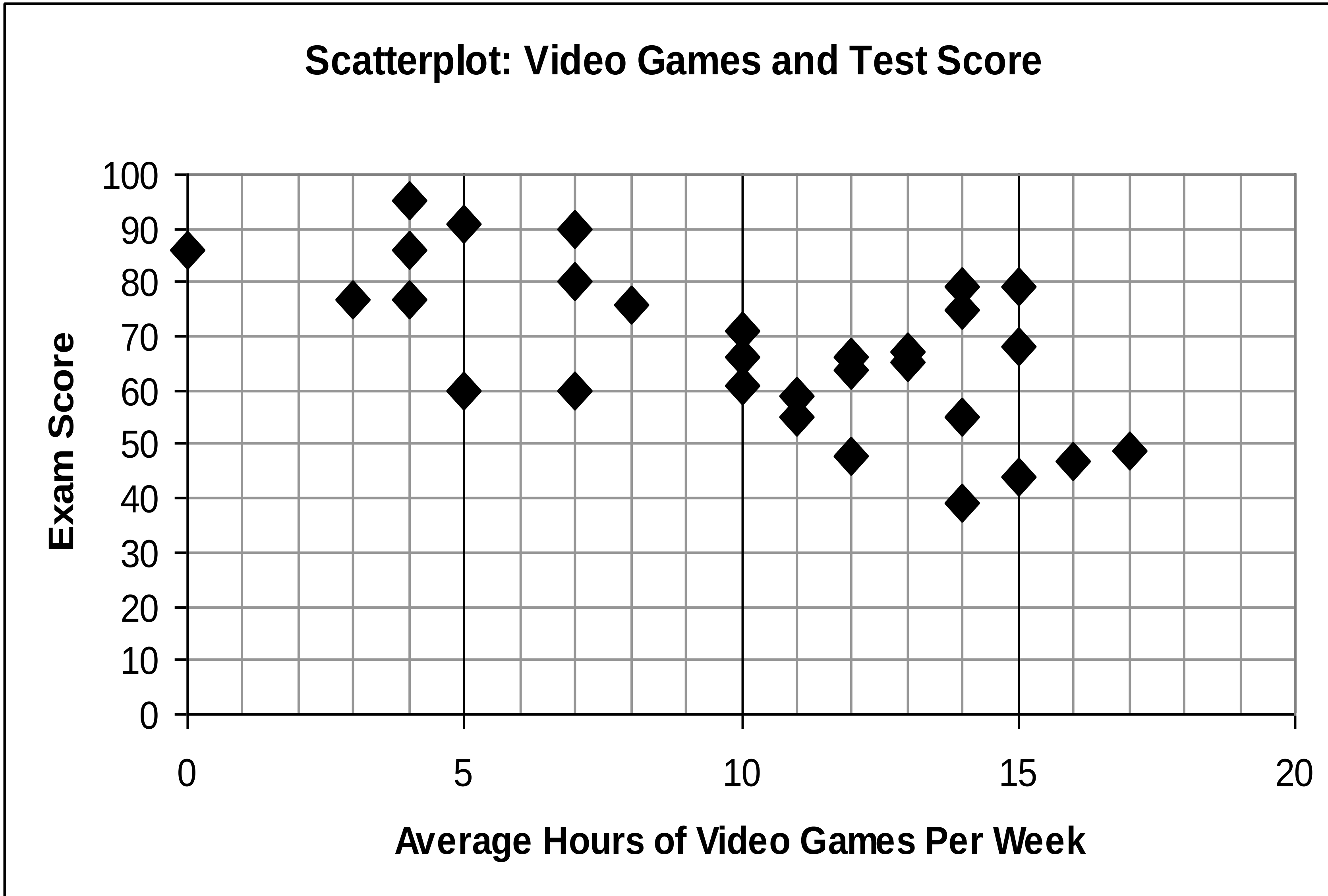
Scatterplots

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

Direct Relationship



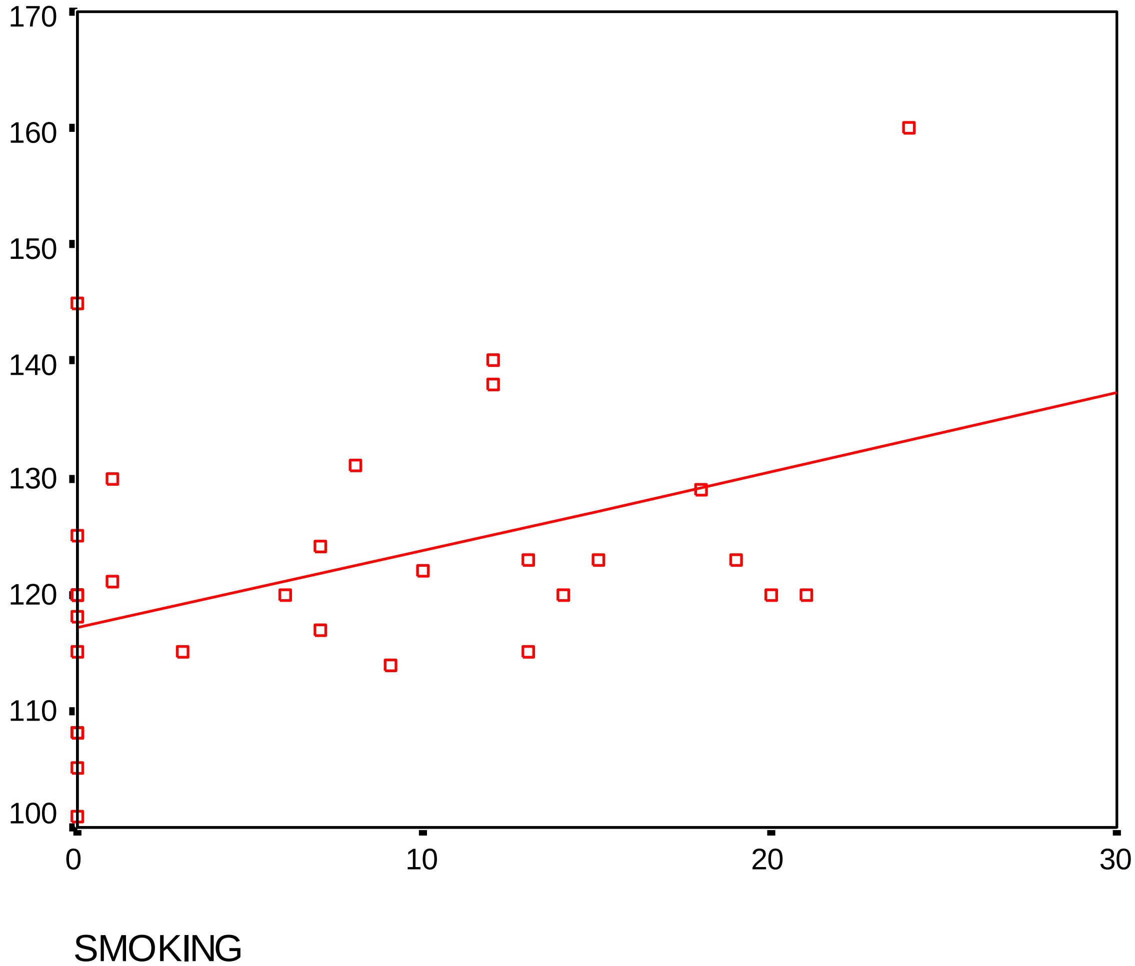
Inverse Relationship



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?



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.

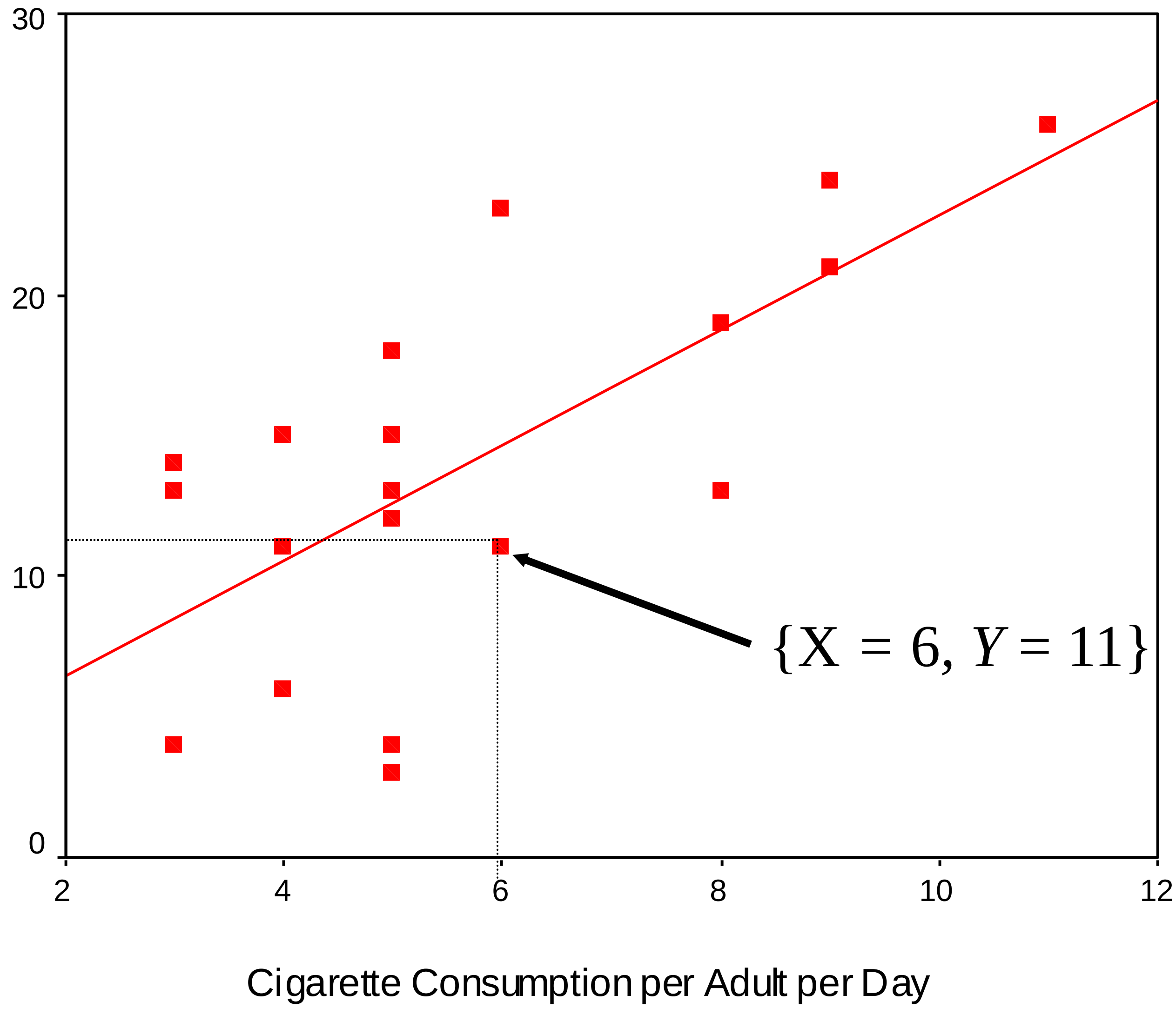
The Data

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
2
4
6
8
10
12
{X = 6, Y = 11}
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

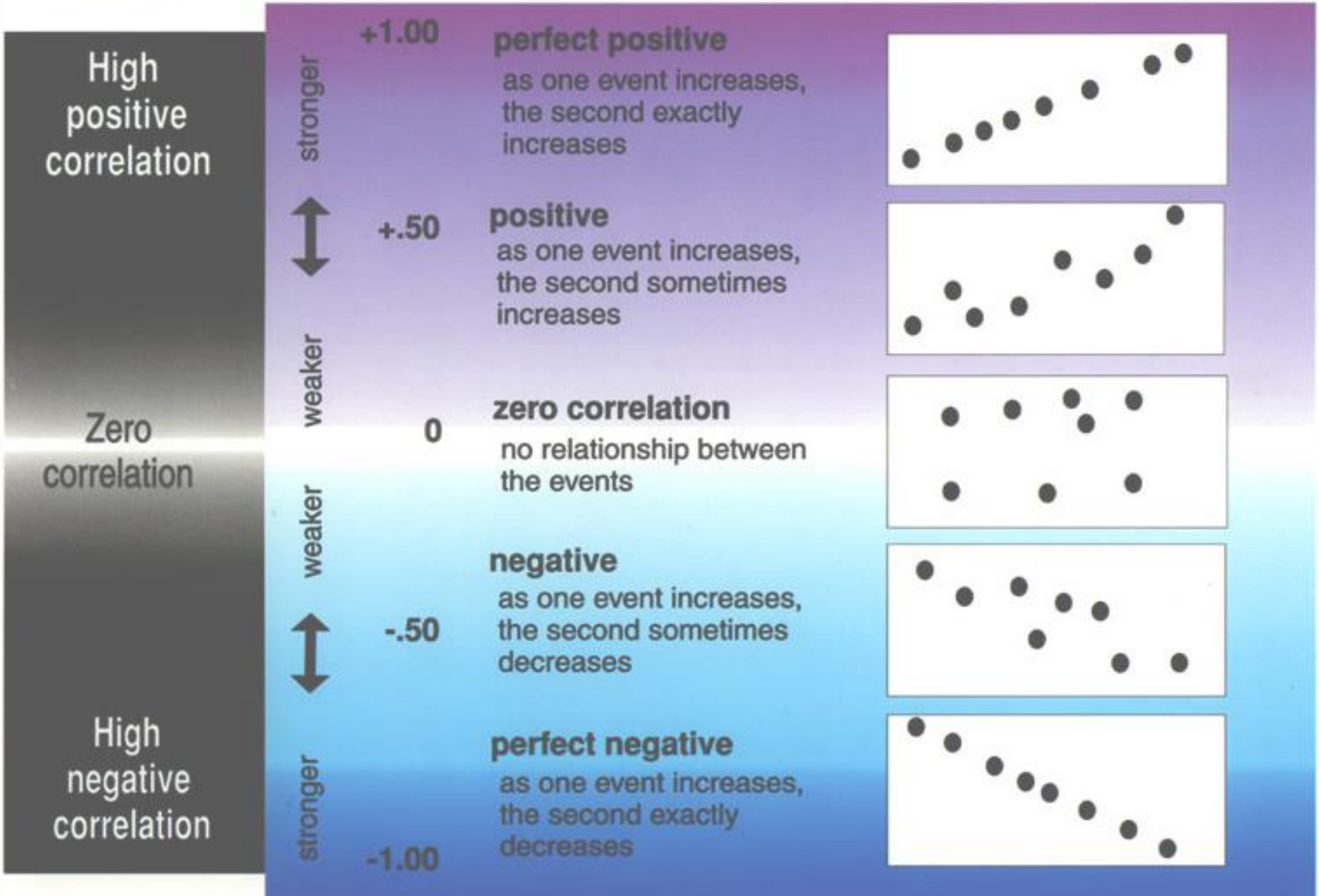
Types of Correlation

- 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

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

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

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

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

Example

- Correlation = .713
- Sign is positive so positive corelation