Correlation

Cal State Northridge
Ψ320
Andrew Ainsworth PhD

Major Points

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

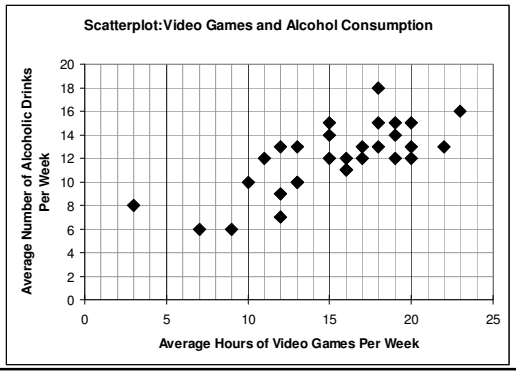
The Question

- 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?

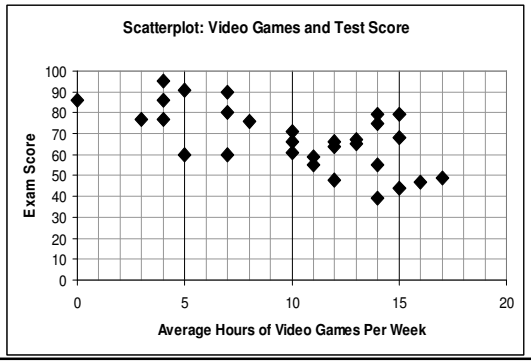
Scatterplots

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

Direct Relationship



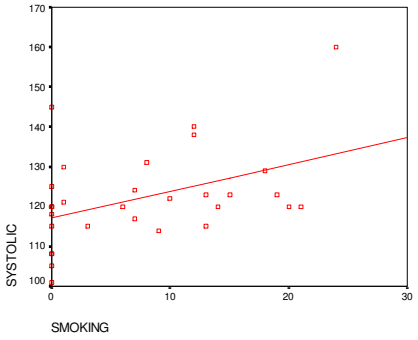
Inverse Relationship



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.

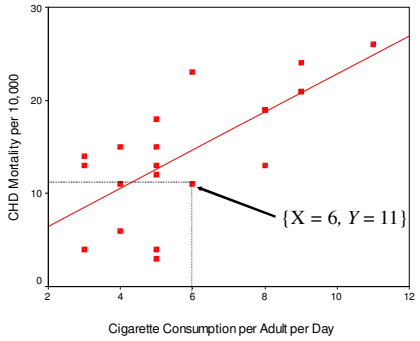
The Data

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



13

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

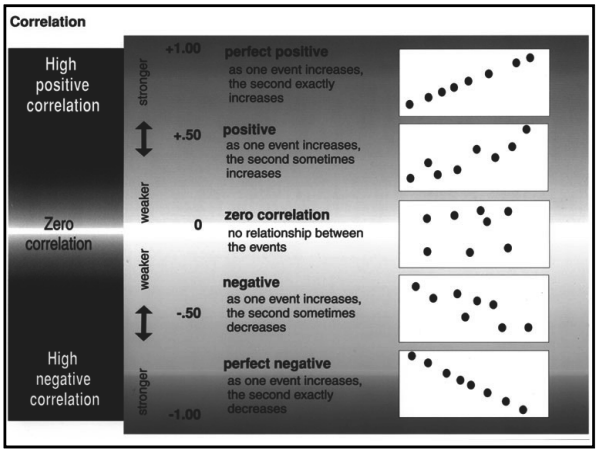
- Co-relation
- The relationship between two variables
- Measured with a correlation coefficient
- Most popularly seen correlation coefficient: Pearson Product-Moment Correlation

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??)



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)	$(X-\overline{X})$	$(Y-\overline{Y})$	$(X-\overline{X})*(Y-\overline{Y})$
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 this was the correlation chapter?

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

- $Cov_{XY} = 11.12$
- $s_X = 2.33$
- $s_Y = 6.69$

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

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

$\Phi = 0.71$

28

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.

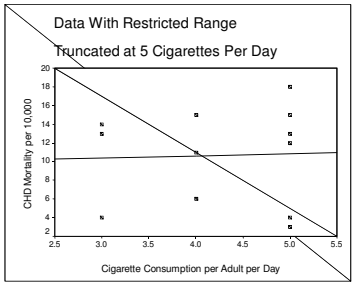
29

Factors Affecting *r*

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

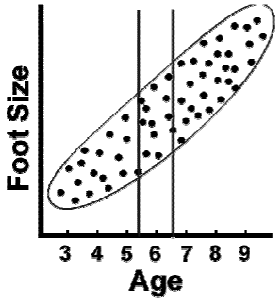
30

Countries With Low Consumptions



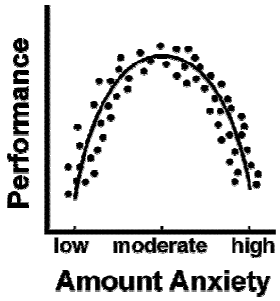
31

Truncation



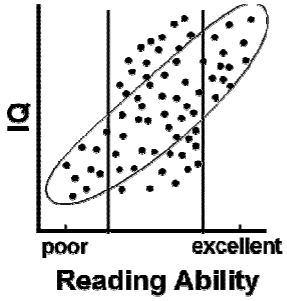
32

Non-linearity



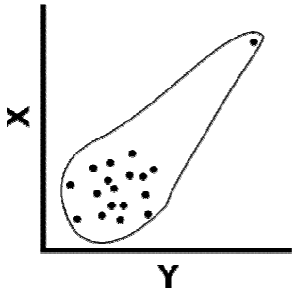
33

Heterogenous samples



34

Outliers



35

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) ρ ≠ 0
  - Two-tailed

Tables of Significance

- Our example *r* was .71
- Table in Appendix E.2
- For *N* - 2 = 19 *df*, r_{crit} = .433
- Our correlation > .433
- Reject H_0
  - Correlation is significant.
  - Greater cigarette consumption associated with higher CHD mortality.

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).