Independent t-test: Compare two means

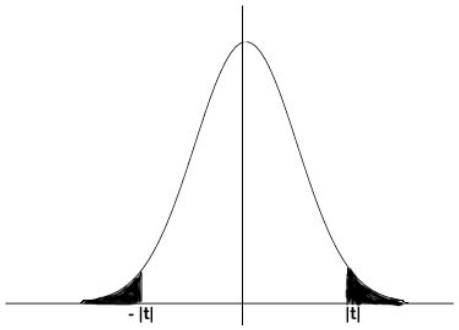
Screen	Productivity Rating
CRT	6.2
CRT	2.7
Flat	5.9
CRT	7.4
Flat	1.5
. . .	. . .

Model (Assumptions) for the independent t-test

- Random sampling, independently from two normal populations
- Possibly different population means
- Same population variance

- Null hypothesis: Population means equal

Two-tailed tests and p-values only!



Robustness of the two-sample t-test

- Normality does not matter much if both samples are large
- Equal variance does not matter much if both samples are large and nearly equal in size
- Independent observations: Important

Matched (paired) t-test

Taste1	Taste2	Difference
10	8	2
7	7	0
3	4	-1
7	8	-1
6	5	1
…	…	…

Within versus between cases

- Between: A case contributes exactly one IV and one DV value
- Within: A case contributes several pairs (IV,DV) - usually one pair for each value of the Independent variable

Model assumptions for matched t-test

- Random sampling of pairs
- Differences are normally distributed (satisfied if both measurements are normal)

Matched t-test

- Null Hypothesis: Mean difference equals zero
- Just a one-sample t-test applied to differences
- Can have more power than an inappropriate independent t-test

Robustness of matched t-test

- For large samples, normality does not matter
- Independent observations matter a lot

One-way analysis of variance

- Could call it “one-factor”
- Could call it “ANOVA”
- Extension of independent t-test: More than two values of the IV
- There are several within-cases versions - not elementary

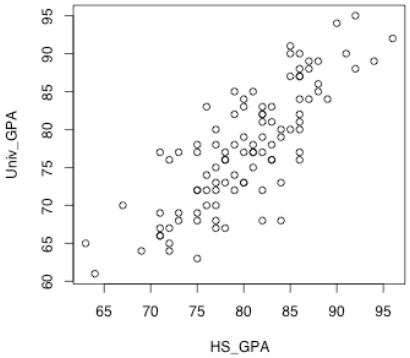
Simple regression and correlation

- Simple means one IV
- DV quantitative
- IV usually quantitative too

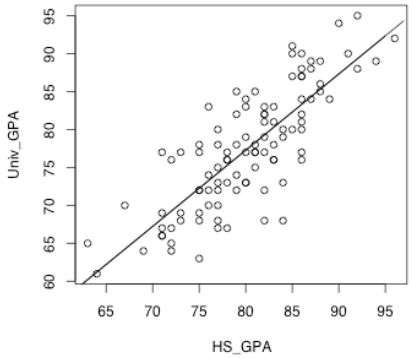
Simple regression and correlation

High School GPA	University GPA
88	86
78	73
87	89
86	81
77	67
…	…

Scatterplot



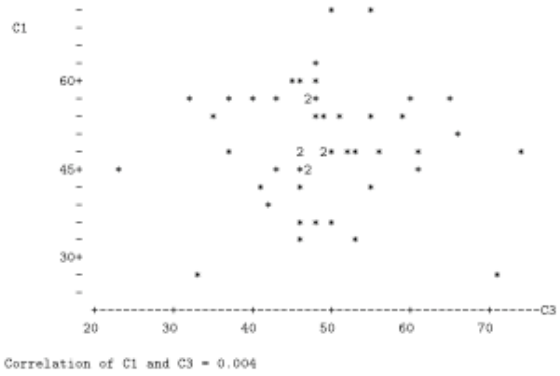
Least squares line



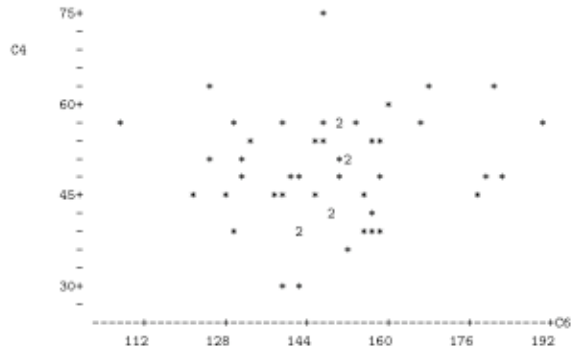
Correlation coefficient *r*

- $-1 \le r \le 1$
- r = +1 indicates a perfect positive linear relationship. All the points are exactly on a line with a positive slope.
- r = -1 indicates a perfect negative linear relationship. All the points are exactly on a line with a negative slope.
- r = 0 means no *linear* relationship (curve possible). Slope of least squares line = 0
- r^2 = proportion of variation explained

r = 0.004

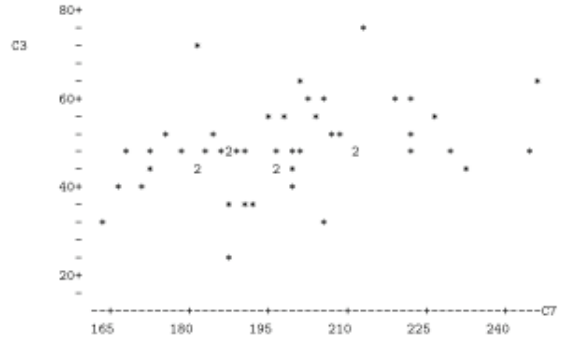


r = 0.112



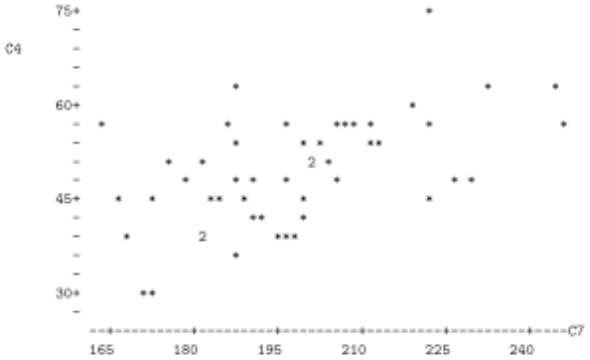
Correlation of C4 and C6 = 0.112

r = 0.368



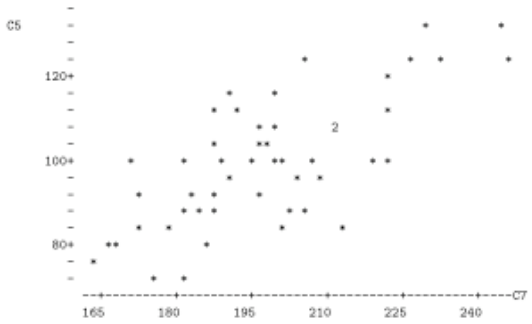
Correlation of C3 and C7 = 0.368

r = 0.547



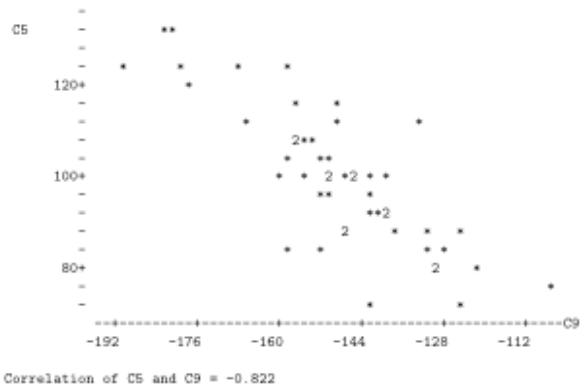
Correlation of C4 and C7 = 0.547

r = 0.733



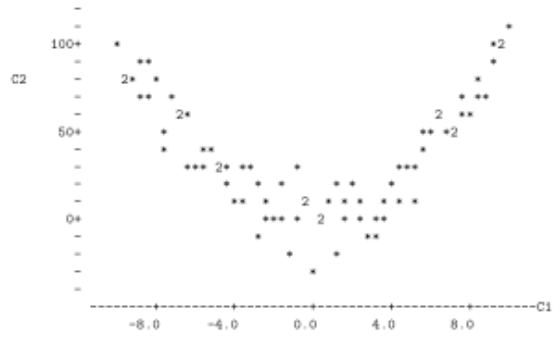
Correlation of C5 and C7 = 0.733

r = - 0.822



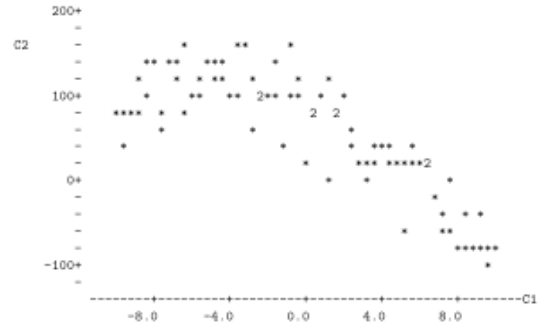
Correlation of C5 and C9 = -0.822

r = 0.025



Correlation of C1 and C2 = 0.025

r = - 0.811



Correlation of C1 and C2 = -0.811

Zero correlation = Horizontal least-squares line

$$\widehat{Y} = b_0 + b_1 X$$

$$b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \overline{Y} - b_1$$

Model assumptions for simple regression

- Random sampling of (X,Y) pairs
- Conditional distribution of DV is normal for each IV value
- Maybe different mean, related to IV by equation of a straight line
- Variances all equal

Testing simple regression

- Null hypothesis: population slope = 0
- (This would make all the conditional distributions identical)
- Same as testing the significance of b_1
- Same as testing the significance of r

Robustness of simple regression test

- Normality does not matter much for large samples if the most influential observations are not too influential.
- Equal variance does not matter much if the number of observations at EACH value of X is large.
- Independent observations: Matters a lot

Chi-square test of independence: Both variables categorical

Music Type	**Stay on Hold?**
A	Yes
A	No
C	Yes
B	Yes
A	No
…	…

"Joint frequency distribution" or "contingency table" or "cross-tabulation" or "crosstab"

	Music Type			
	A	**B**	**C**	**D**
Yes	41	15	38	45
No	9	35	12	5

Model assumptions for the chi-squared test of independence

- The variable consisting of combinations of IV, DV has a multinomial distribution
- "Large" random sample
- Rule of thumb: Lowest expected frequency no more than 5

- Independent observations: Important and often violated in practice.

Formula for the chi-square test

$$\chi^2 = \sum_{\text{cells}} \frac{(f_o - f_e)^2}{f_e}$$

- Even one very small expected frequency can make chisquare huge
- Smallest expected frequency no more than one (not 5) controls Type I error

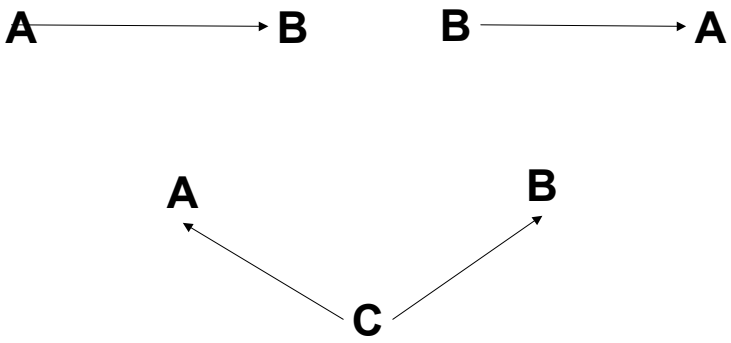
Why predict DV from IV?

- There may be a practical reason for prediction (buy, make a claim, price of wheat).
- It may be "science."

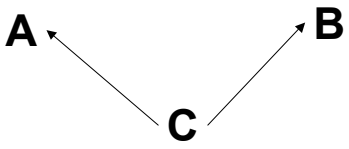
Young smokers who buy contraband cigarettes tend to smoke more.

- What is IV, DV?

Correlation is not the same as causation



Confounding variable: A variable that contributes to both IV and DV, causing a misleading relationship between them.



Mozart Effect

- Babies who listen to classical music tend to do better in school later on.
- Does this mean parents should play classical music for their babies?
- Please comment. (What is one possible confounding variable?)

Hypothetical study

- Subjects are babies in an orphanage (maybe in Haiti) awaiting adoption in Canada. All are assigned, but waiting for the paperwork to clear.
- They all wear headphones 5 hours a day. Randomly assigned to classical, rock, hip-hop or nature sounds. Same volume.
- Assess academic progress in JK, SJ, Grade 4.
- Suppose there is a significant difference? What are some potential confounding variables?

Experimental vs. Observational studies

- **Observational**: IV, DV just observed and recorded
- **Experimental**: Cases randomly assigned to values of IV
- Only a true experimental study can establish a causal connection between IV and DV

- Maybe we should talk about observational vs experimental variables.
- Watch it: Confounding variables can creep back in.