

# STA 441: Data Analysis

**This slide show is an open-source document.  
See last slide for copyright information.**

# Data Science

- Study design
- Data acquisition
- Data processing and perhaps pre-cleaning, yielding a data file.
- Data cleaning and description
- Data analysis and usually more cleaning.
- Interpretation, possibly with recommendations.
- Action

# Data Science

- Study design
- Data acquisition
- Data processing and perhaps pre-cleaning, yielding a data file.
- Data cleaning and description
- Data analysis and usually more cleaning.
- Interpretation, possibly with recommendations.
- Action

# Data File

- Rows are **cases**
- Columns are **variables**

1	2	2	0	78.0	65	80	39	English	Female	3	3	1
2	2	6	2	66.0	54	75	57	English	Female	3	3	1
3	2	4	4	80.2	77	70	62	English	Male	5	6	1
4	2	5	2	81.7	80	67	76	English	Female	2	2	1
5	2	4	4	86.8	87	80	86	English	Male	5	5	1
6	2	3	1	76.7	53	75	60	English	Male	3	3	1
7	2	3	2	85.8	86	81	54	Other	Female	2	2	1
8	2	4	3	73.0	75	77	17	English	Male	4	5	1
9	2	6	2	72.3	63	60	2	English	Male	4	4	1
10	2	8	6	90.3	87	88	76	English	Male	4	4	1
11	2	8	3	.	.	.	60	English	Male	1	2	1
12	2	6	4	.	.	.	61	Other	Female	1	1	1
13	.	.	.	87.2	84	83	54	English	Male	3	3	1
14	2	2	5	91.0	90	91	84	English	Male	5	5	1
15	2	3	1	72.8	53	74	.	English	Female	3	3	1
16	.	.	.	80.7	72	84	14	English	Male	3	3	1
17	2	5	0	82.5	82	85	75	Other	Female	2	2	1
18	2	4	6	91.5	95	81	94	English	Female	3	3	1
19	2	3	2	78.3	77	74	60	English	Female	3	3	1
20	.	.	.	74.5	0	85	.	English	Male	4	4	1
21	2	3	3	80.7	71	78	53	Other	Female	1	3	1
22	2	5	3	88.3	80	85	63	English	Female	3	3	1
23	2	4	2	76.8	82	64	82	Other	Female	2	2	1

Skipping ....

570	2	5	4	84.8	88	68	80	English	Male	1	1	1
571	2	4	3	78.3	83	84	56	English	Male	4	2	1
572	2	6	3	88.3	81	90	70	English	Female	5	5	1
573	2	3	1	.	.	.	.	English	Male	3	3	1
574	2	5	9	77.0	73	79	60	English	Female	2	2	1
575	.	.	.	78.7	80	73	.	English	Female	6	3	1
576	2	5	2	80.7	80	70	50	Other	Male	1	1	1
577	2	4	2	80.7	56	81	50	English	Female	2	2	1
578	2	4	3	.	.	.	78	Other	Female	4	4	1
579	1	6	1	82.2	80	86	61	English	Female	2	2	1

id	mcg	r	day	AML	AMS	AMld	PML	PMS	PMld	AMslp	PMslp	SWeight
1	198	1	1	0.6	.	.	0.8	.	.	.	.	.
2	198	1	2	1.8	.	.	2.8	.	.	.	.	.
3	198	1	3	4.7	1	.	6.1	1	.	.	.	.
4	198	1	4	7.8	4	2.0	8.7	5	2.1	.	.	.
5	198	1	5	11.2	6	1.8	12.1	7	2.0	.	.	.
6	198	1	6	14.3	12	1.9	15.0	11	1.4	.	.	.
7	198	1	7	17.5	12	2.1	18.5	13	1.6	.	.	.
8	198	1	8	20.9	19	1.1	21.9	19	1.7	.	.	.
9	198	1	9	24.0	22	1.6	25.2	22	1.3	.	.	.
10	198	1	10	27.2	26	2.1	28.4	26	1.2	.	.	.
11	198	1	11	30.7	28	1.4	32.3	28	1.5	.	.	.
12	198	1	12	.	31	.	.	31	.	.	.	.
13	198	1	13	.	37	.	.	36	.	.	.	.
14	198	1	14	.	37	.	.	38	.	3.11	3.18	0.5996
15	198	2	1	0.5	.	.	0.6	.	.	.	.	.
16	198	2	2	1.4	.	.	2.3	.	.	.	.	.
17	198	2	3	4.15	1	.	5.6	1	.	.	.	.
18	198	2	4	7.4	2	2.0	8.7	4	2.1	.	.	.
19	198	2	5	10.8	5	2.2	12.0	8	2.0	.	.	.
20	198	2	6	14.2	10	1.7	15.3	13	1.6	.	.	.
21	198	2	7	17.1	13	2.2	18.1	16	1.7	.	.	.
22	198	2	8	21.3	18	1.1	22.2	18	1.4	.	.	.
23	198	2	9	24.4	27	1.4	25.6	24	1.2	.	.	.
24	198	2	10	27.6	26	2.1	28.8	28	1.2	.	.	.
25	198	2	11	31.2	29	1.9	32.5	29	1.3	.	.	.
26	198	2	12	.	33	.	.	36	.	.	.	.
27	198	2	13	.	38	.	.	41	.	.	.	.
28	198	2	14	.	42	.	.	42	.	3.21	3.26	0.6040

# Variables can be

- Quantitative - representing amount of something, like Income, BP, BMI, GPA (?)
- Categorical - Codes represent category membership, like Gender, Nationality, Marital status, Alive vs. dead

# Variables can be

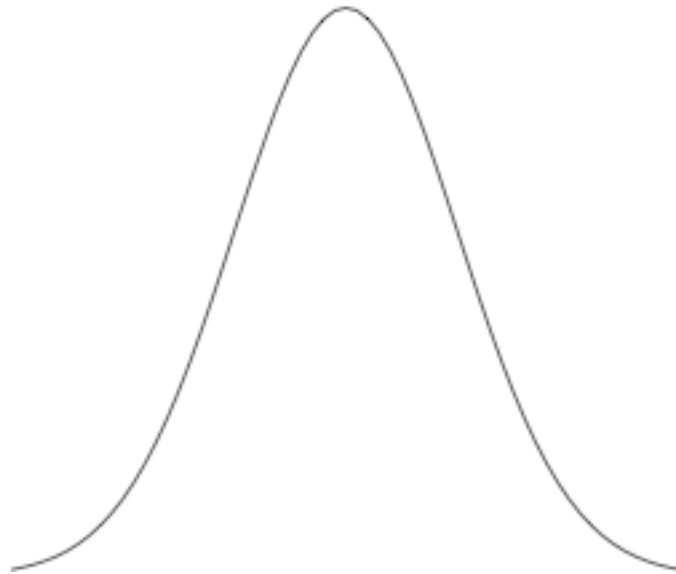
- Explanatory: Predictor or cause (contributing factor)
- Response: Predicted or effect



We will often pretend that our data represent a **random sample** from some **population**. We will carry out formal procedures for making inferences about this (usually fictitious) population, and then use them as a basis for drawing conclusions about the data.

- **Statistics:** Numbers that can be calculated from sample data
- **Parameters:** Numbers that could be calculated if we knew the whole population

# **Distribution = Population Histogram**



# Conditional Distribution

For each value  $x$  of the explanatory variable  $X$ , there is a separate distribution of the response Variable  $Y$ . This is called the conditional distribution of  $Y$  given  $X=x$ .

Example: Conditional distribution of height given  
Gender = F.

# Definition of “Related”

- We will say that the explanatory and response variables are **unrelated** if the conditional distribution of the response variable is identical for each value of the explanatory variable.
- If the distribution of the response variable does depend on the value of the explanatory variable, we will describe the two variables as **related**.

# Testing Statistical Significance

- Are explanatory variable and response variable “really” related?
- **Null Hypothesis:** They are unrelated in the population.

# Reasoning

Suppose that the explanatory and response variables are actually unrelated in the population. If this null hypothesis is true, what is the probability of obtaining a sample relationship between the variables that is as strong or stronger than the one we have observed? If the probability is small (say,  $p < 0.05$ ), then we describe the sample relationship as **statistically significant**, and it is socially acceptable to discuss the results.

# P-value

- The probability of getting our results (or better) just by chance.
- The minimum significance level at which the null hypothesis can be rejected.



# We can be wrong

- Type I error:  $H_0$  is true, but we reject it
- Type II error:  $H_0$  is false, but we fail to reject it

# **Power** is the probability of *correctly* rejecting $H_0$

- Power =  $1 - P(\text{Type II Error})$
- Power increases with true strength of relationship, and with sample size
- Power can be used to select sample size in advance of data collection

**Confidence Interval:** Pair of numbers chosen so that the probability they will enclose the parameter (or function of parameters) is large, like 0.95

# From the CBC Website

The Angus Reid Institute said its online survey was conducted among a representative randomized sample of 3,749 Canadian adults who are members of the Angus Reid Forum. For comparison purposes, a probability sample of this size would carry a margin of error of plus or minus 1.5 percentage points, 19 times out of 20.

# Should we Accept $H_0$ ?

- This is a philosophic question.
- A historic debate/fight between Neyman and Pearson on one side, and R. A. F. Fisher on the other side.
- In this class, we will agree with Fisher, and the answer is **NO**.

# What to say if results are not statistically significant

- Do **not** say there was no difference, or no relationship – half marks off.
- Try not to mention the results at all.
- If you are directly asked, or if the test is intended to answer a central question of the study, you could say that “the data do not provide enough evidence to conclude that the variables are related.”
- Or, ...

# More ways to say nothing happened

- There was no meaningful connection between ...
- Results were consistent with no treatment effect.
- Results were consistent with no association between astrological sign and personality type.
- The small differences in average taste ratings could have been due to sampling error.
- The small differences in average taste ratings were within the range of sampling error.
- Preference for the two brands was in a statistical tie.

**The nice thing about using this kind of language is that it communicates clearly to non-experts, but it lets the experts read between the lines and see that you are aware of the technical (philosophic) issue, and that you are being careful.**

# Many statistical methods assume **Independent Observations**

- Simple random sampling
- Cases are not linked, do not “communicate”
- If the design involves non-independence, allow for it.



# Elementary Tests

- Independent (two-sample) t-test
- Matched (paired) t-test
- One-way ANOVA
- Simple regression and correlation
- Chi-square test of independence

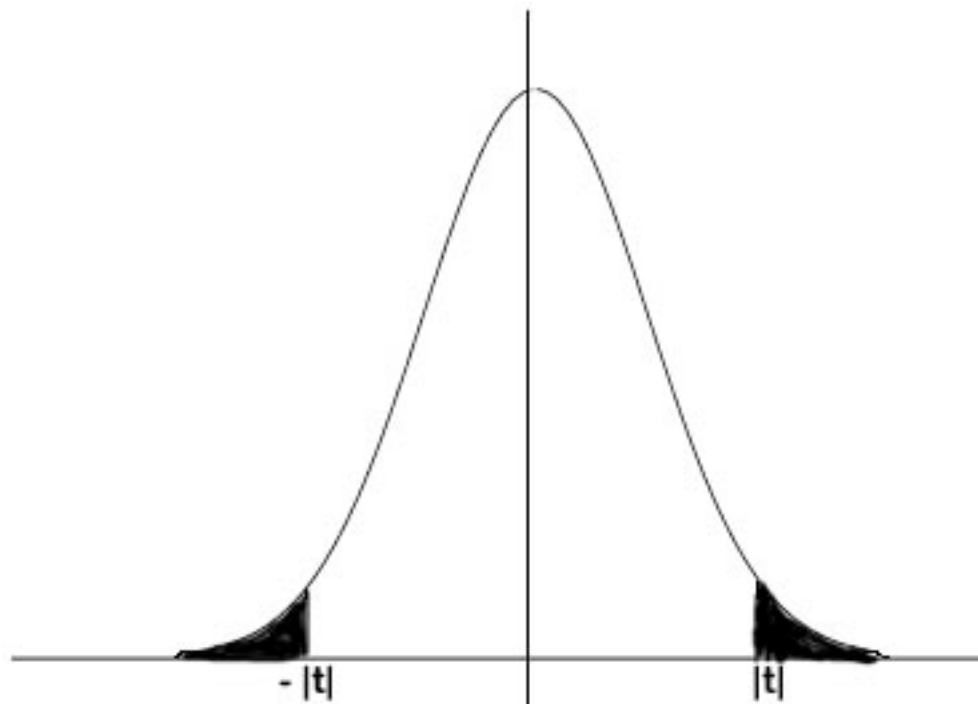
# Independent t-test: Compare two means

<b>Data Plan</b>	<b>Productivity Rating</b>
A	6.2
A	2.7
B	5.9
A	7.4
B	1.5
...	...

# Model (Assumptions) for the independent t-test

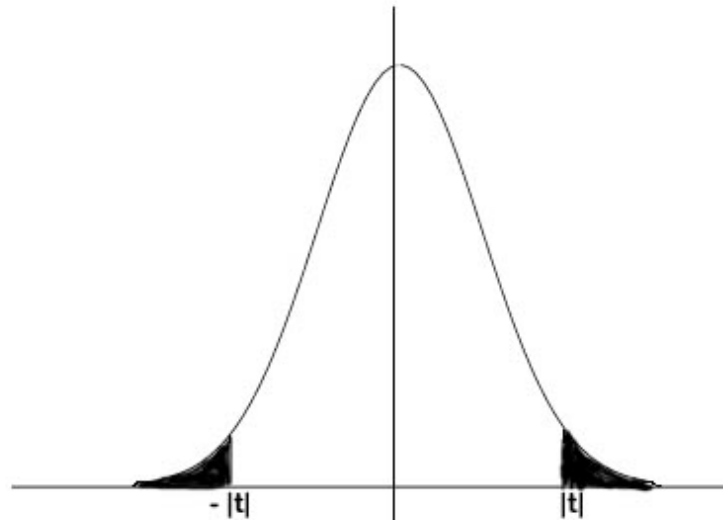
- Random sampling, independently from two normal populations
- Possibly different population means
- Same population variance
- Null hypothesis: Population means equal

# Two-tailed tests and p-values only!



But we will always draw directional conclusions  
when possible

- Look at the sign of the regression coefficient
- Look at the sample means
- Look at the sample percentages



# 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 explanatory variable and one response variable value
- Within: A case contributes several pairs (explanatory variable, response variable) - usually one pair for each value of the explanatory 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 explanatory variable
- There are several within-cases versions  
- not elementary

# Simple regression and correlation

- Simple means one explanatory variable
- Response variable quantitative
- Explanatory variable usually quantitative too

# Simple regression and correlation

**High School GPA**

**University GPA**

88

86

78

73

87

89

86

81

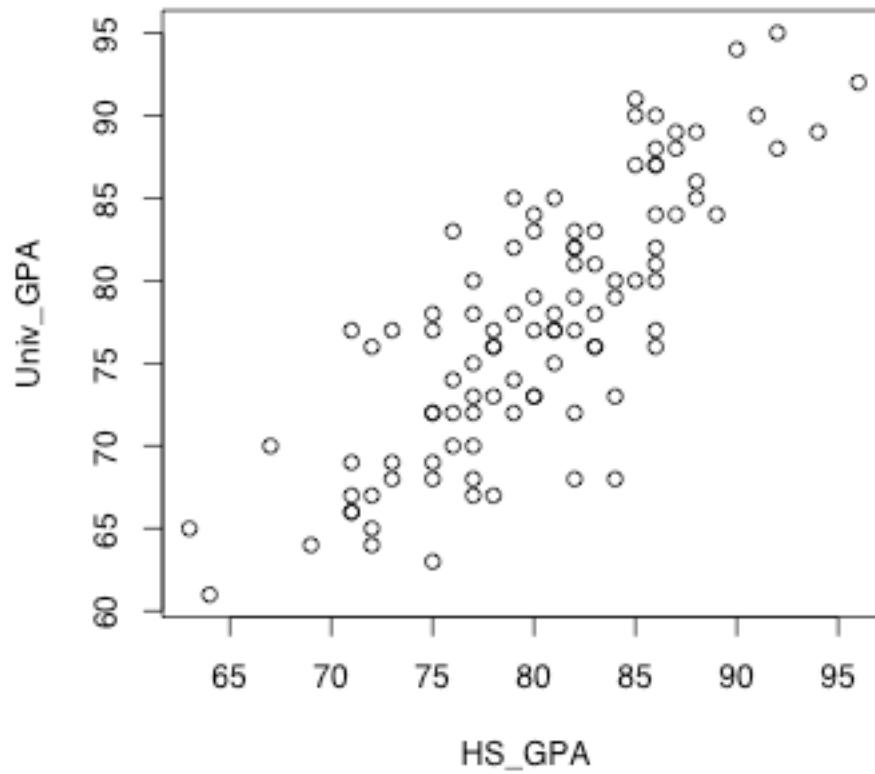
77

67

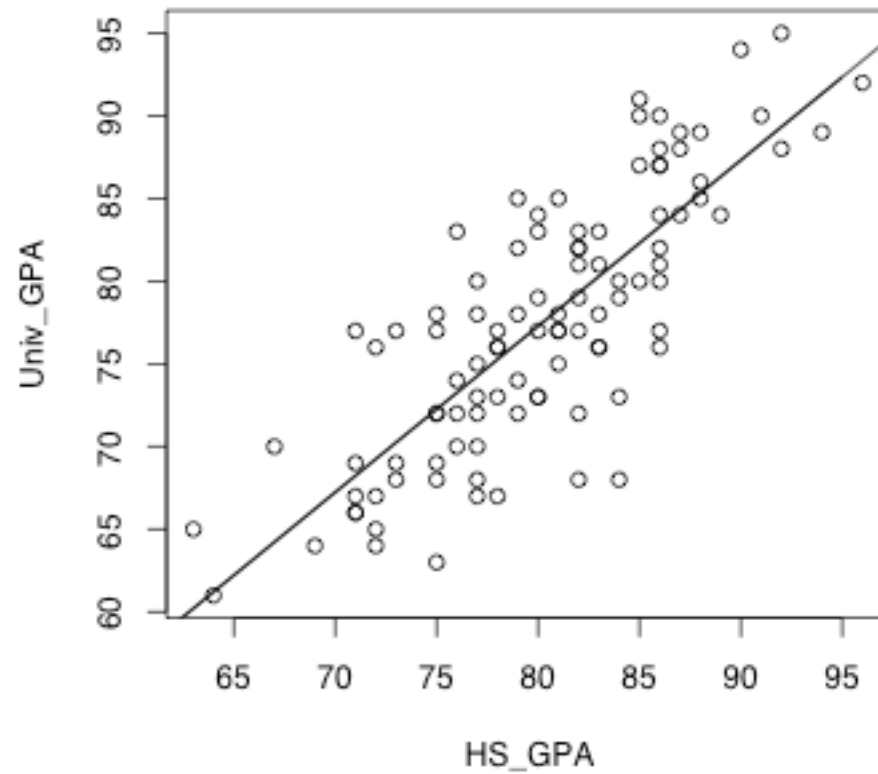
...

...

# Scatterplot



# Least squares line

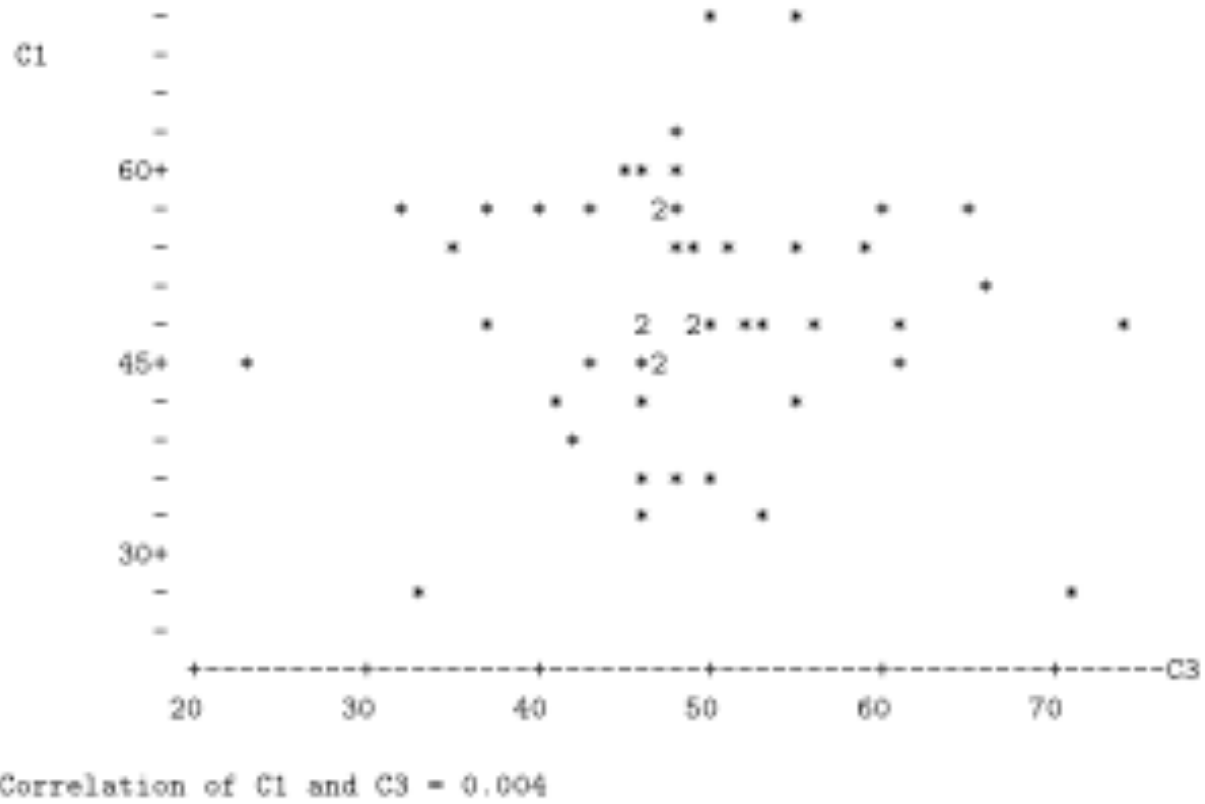




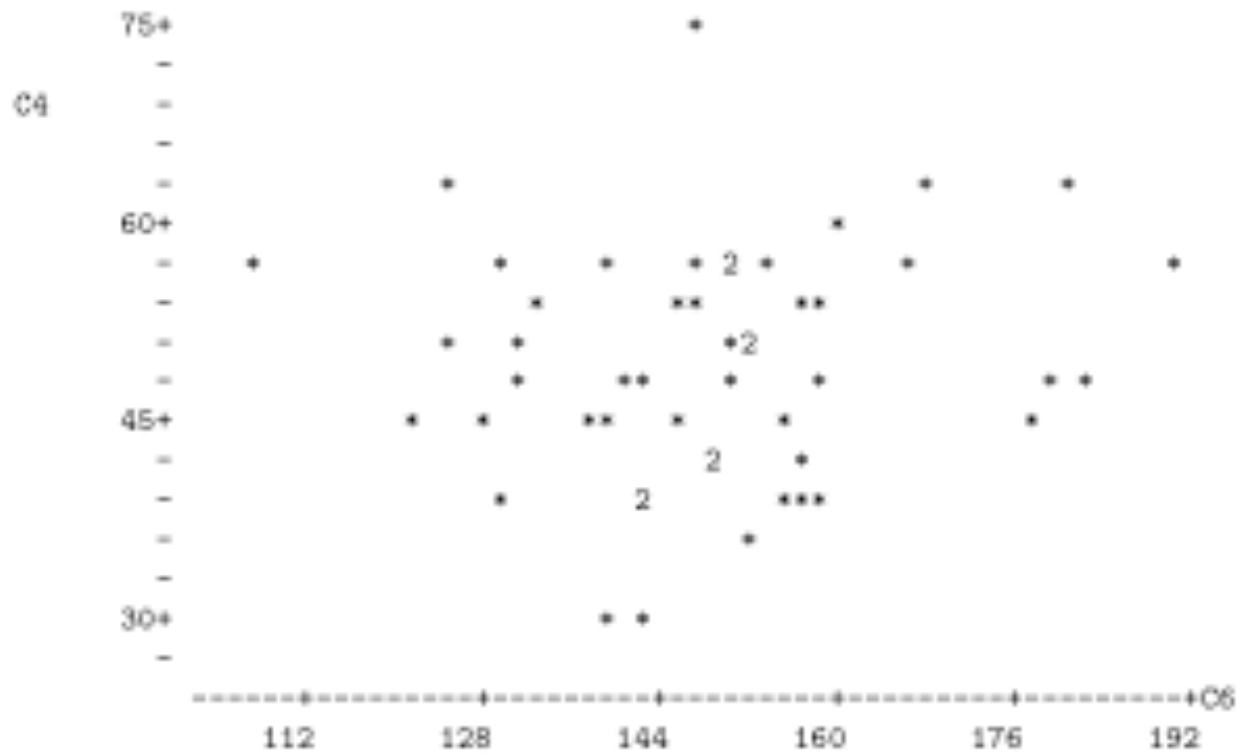
# Correlation coefficient $r$

- $-1 \leq r \leq 1$
- $r = +1$  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

$r = 0.004$

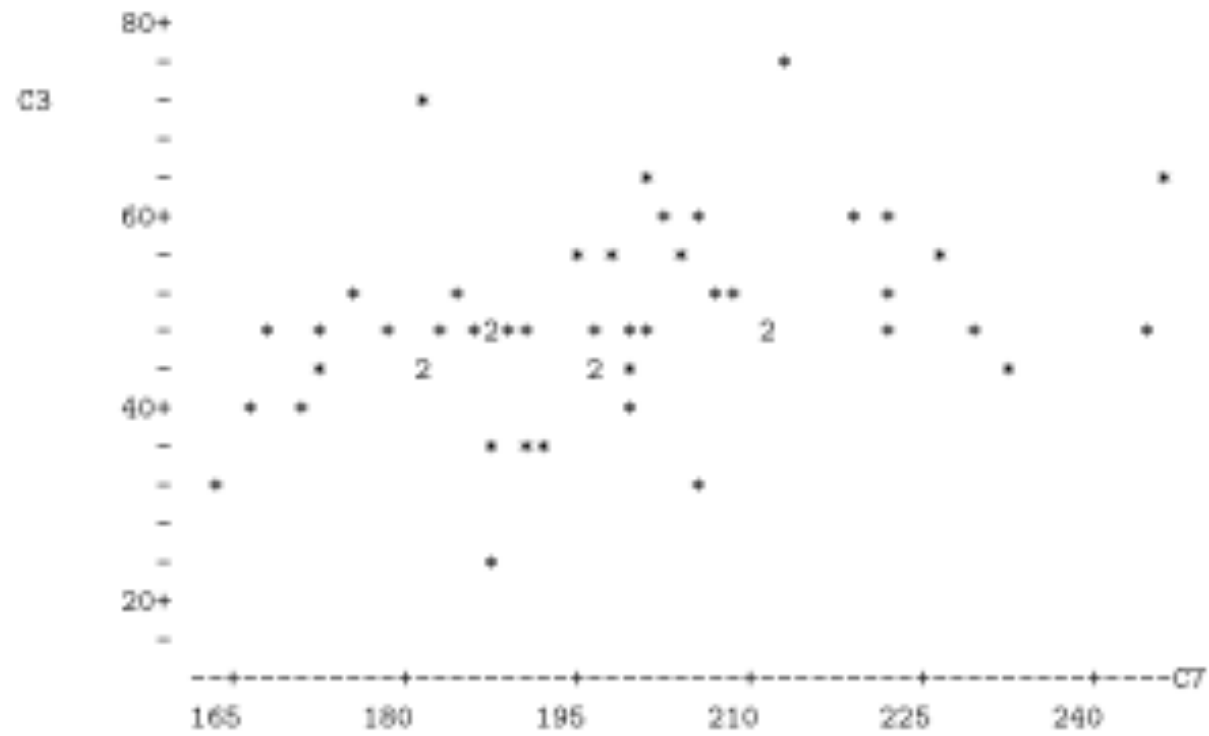


$r = 0.112$



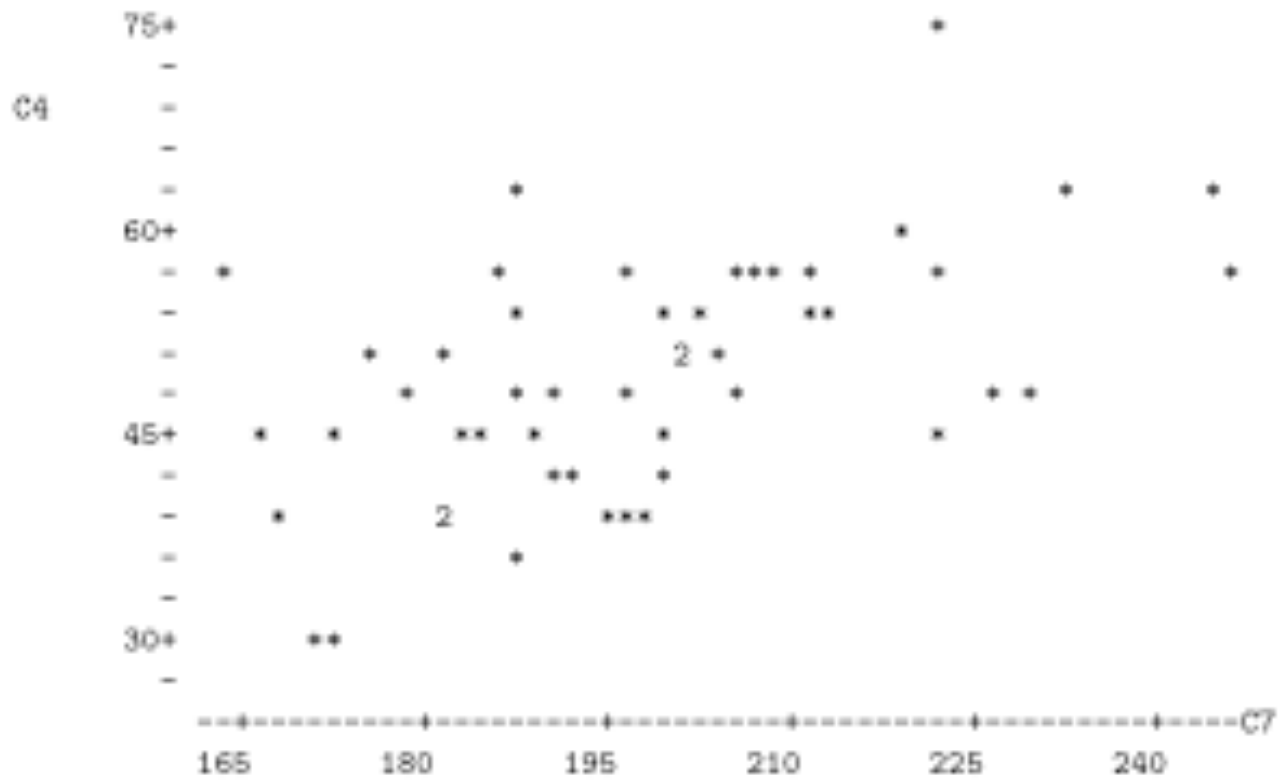
Correlation of C4 and C6 = 0.112

$$r = 0.368$$



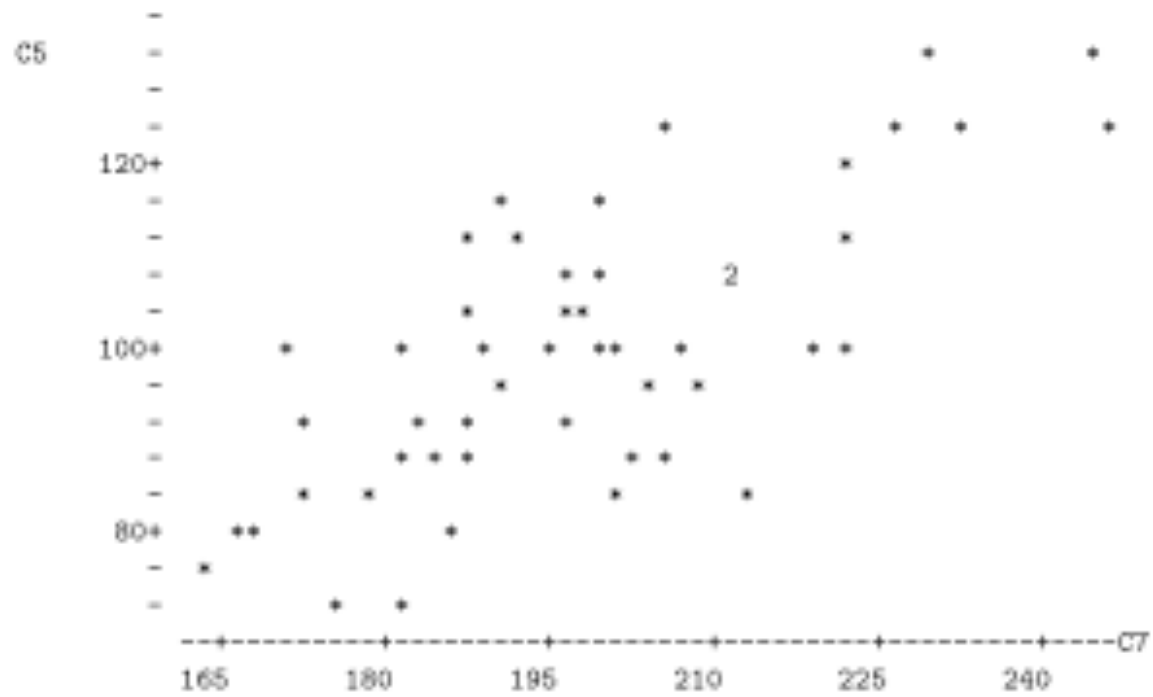
Correlation of C3 and C7 = 0.368

$$r = 0.547$$



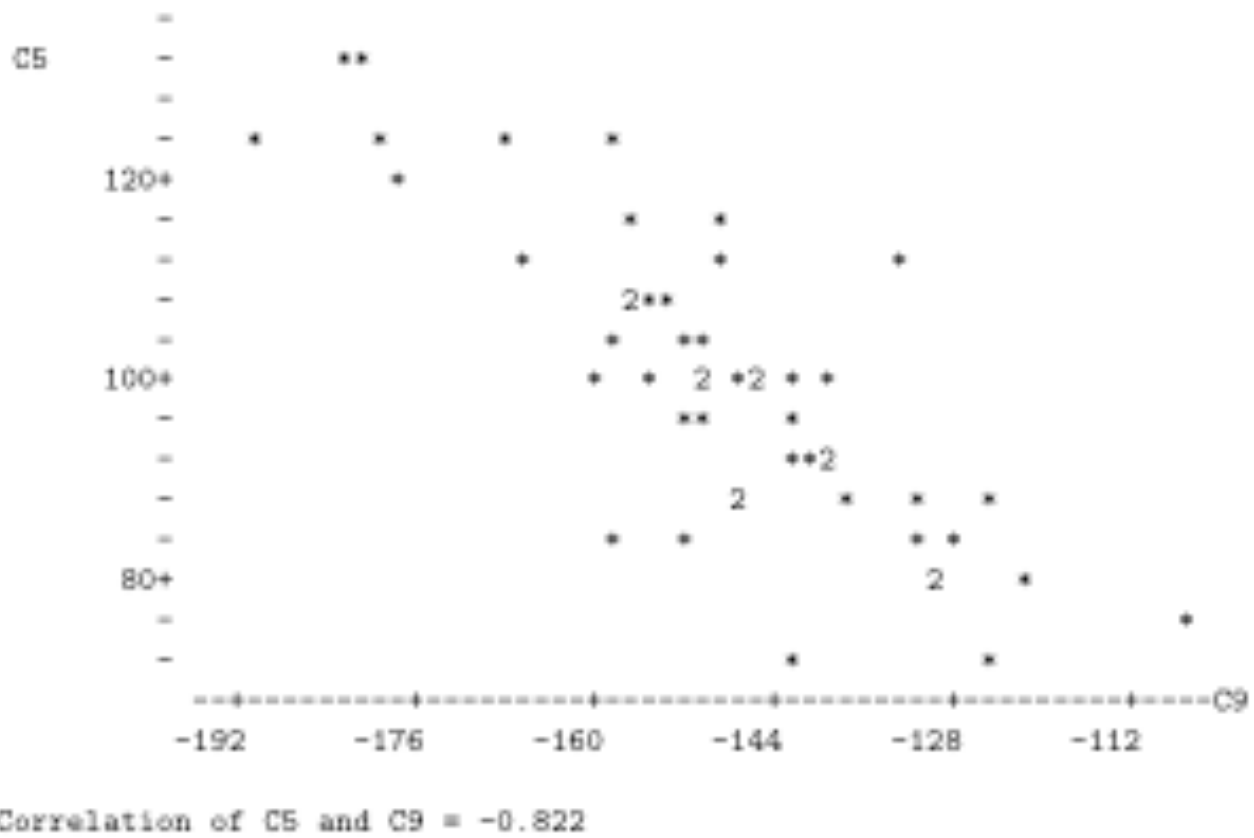
Correlation of C4 and C7 = 0.547

$$r = 0.733$$

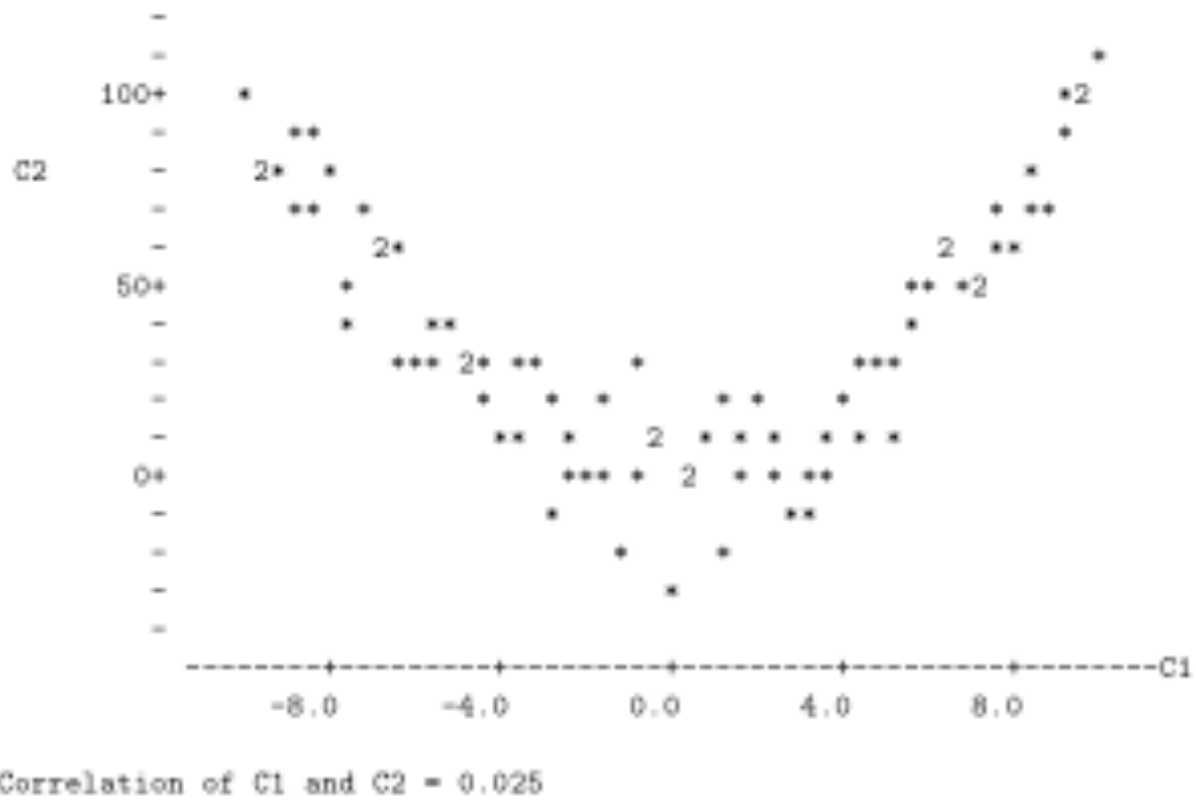


Correlation of C5 and C7 = 0.733

$$r = -0.822$$

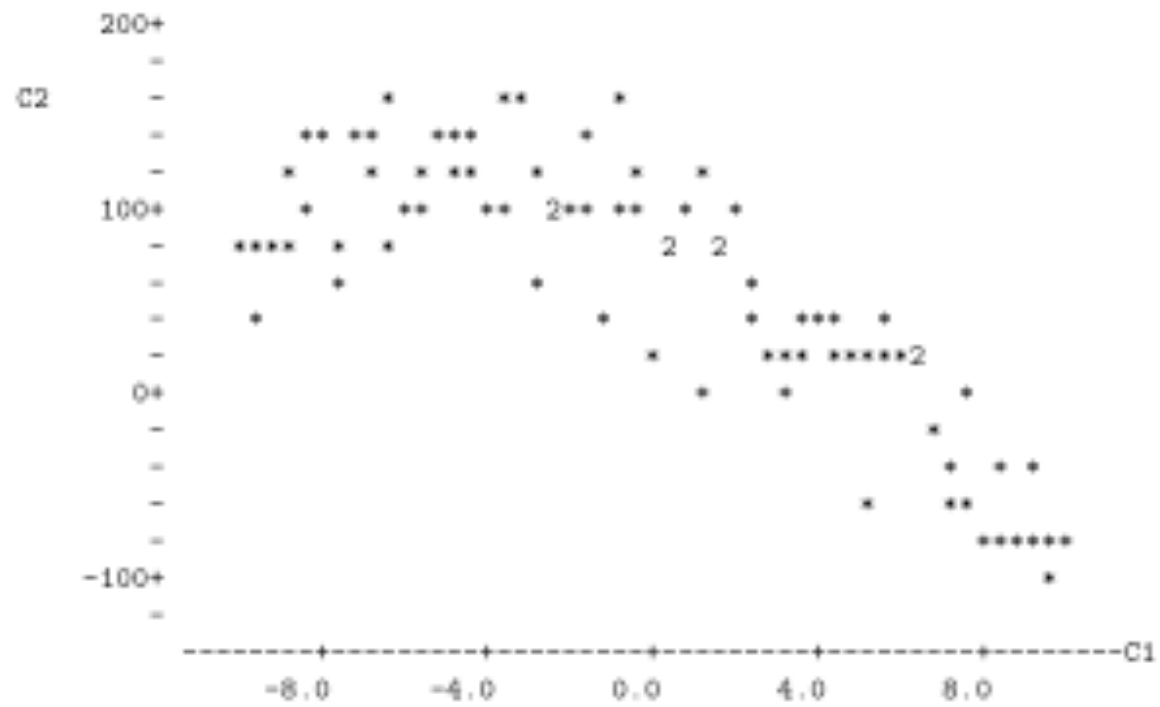


$$r = 0.025$$





$$r = -0.811$$



Correlation of C1 and C2 = -0.811

Zero correlation = Horizontal  
least-squares line

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

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

# Model assumptions for simple regression

- Random sampling of  $(X, Y)$  pairs
- Conditional distribution of response variable is normal for each explanatory variable value
- Maybe different mean, related to explanatory variable 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-squared test of independence: Both variables categorical

<b>Music Type</b>	<b>Stay on Hold?</b>
A	Yes
A	No
C	Yes
B	Yes
A	No
...	...

“Joint frequency distribution” or  
“contingency table” or “cross-  
tabulation” or “crosstab”

	Music Type			
	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>
<b>Yes</b>	41	15	38	45
<b>No</b>	9	35	12	5

## Model assumptions for the chi-squared test of independence

- The variable consisting of combinations of explanatory variable, response variable 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-squared test

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

- Even one very small expected frequency can make the chisquared statistic huge.
- Smallest expected frequency no less than one (not 5) controls Type I error okay.

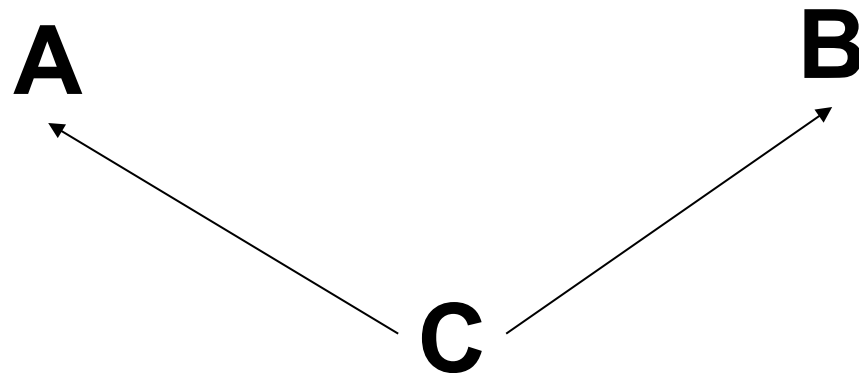
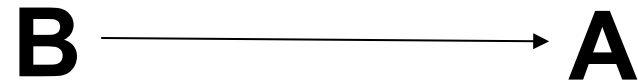
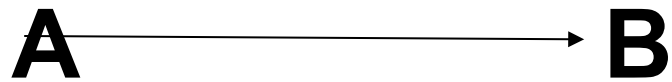
# Why predict response variable from explanatory variable?

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

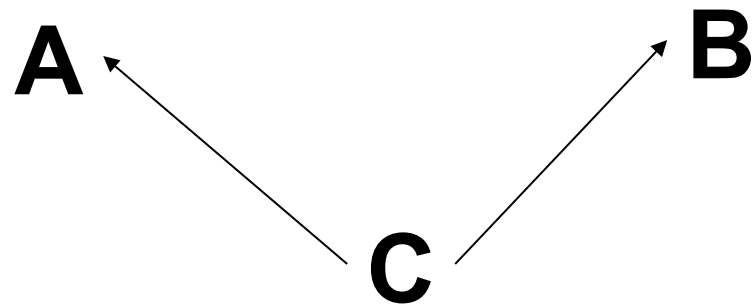
Young smokers who buy contraband cigarettes tend to smoke more.

- What is explanatory variable, response variable?

# Correlation is not the same as causation



**Confounding variable:** A variable that contributes to both explanatory variable and response variable, 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 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.
- Adoptive parents not informed.
- 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:** explanatory variable, response variable just observed and recorded
- **Experimental:** Cases randomly assigned to values of explanatory variable
- Only a true experimental study can establish a causal connection between explanatory variable and response variable
  
- Maybe we should talk about observational vs experimental variables.
- Watch it: Confounding variables can creep back in.



# Marking rule

- If you are interpreting the results of a purely observational study and you state an unqualified causal connection between explanatory and response variable, you lose half marks on the interpretation part of the question.
- Examples:
  - Exercise reduces arthritis pain.
  - Higher doses of Vitamin C lead to fewer colds.
  - Higher income produces greater average reported happiness.
  - More interaction with co-workers leads to greater job satisfaction.
  - Textbook had a large effect.
  - Religion influences number of children.

# Plain language is important

- If you can only be understood by mathematicians and statisticians, your knowledge is much less valuable.
- Often a question will say “Give the answer in plain, non-statistical language.”
- This means if  $x$  is income and  $y$  is credit card debt, you make a statement about income and average or predicted credit card debt, like “Customers with higher incomes tend to have less credit card debt.”
- If you use mathematical notation or words like null hypothesis, unbiased estimator, p-value or statistically significant, you will lose a lot of marks even if the statement is correct. Even avoid “positive relationship,” and so on.

# Plain language

- If the study is about fish, talk about fish.
- If the study is about blood pressure, talk about blood pressure.
- If the study is about breaking strength of yarn, talk about breaking strength of yarn.
- Assume you are talking to your boss, a former Commerce major who got a D+ in ECO220 and does not like to feel stupid.

# We will be guided by tests with $\alpha = 0.05$

- If a null hypothesis like  $H_0: \beta_1=0$  is rejected, we will draw a conclusion.
- If we do not reject the null hypothesis, we will not draw a definite conclusion.
- Instead, say things like:
  - There is no evidence of a connection between blood sugar level and mood.
  - These results are not strong enough for us to conclude that attractiveness is related to mark in first-year Computer Science.
  - These results are consistent with no effect of dosage level on bone density.
- Though we are guided by  $\alpha = 0.05$ , we *never* mention it when plain language is required.

# No one-tailed tests

- In this class we will avoid one-tailed tests.
- Why? Ask what would happen if the results were strong and in the opposite direction to what was predicted.
- If the question asks for a null hypothesis and your answer has an inequality, it's wrong.
- But when  $H_0$  is rejected, we still draw directional conclusions.

# Directional conclusions

- Suppose  $x$  is income and  $y$  is credit card debt, and we test  $H_0: \beta_1=0$  with a two-sided t-test.
- Say  $p = 0.0021$  and  $b_1 = 1.27$ .
- We say “Consumers with higher incomes tend to have more credit card debt.”
- Is this justified? We'd better hope so, or all we can say is “There is a connection between income and average credit card debt.”
- Then they ask: “What's the connection? Do people with lower income have more debt?”
- And you have to say “Sorry, I don't know.”
- It's a good way to get fired, or at least look silly.
- If a directional conclusion is possible and you only say “related,” you get half marks **at most**.

# Copyright Information

This slide show was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. These Powerpoint slides are available from the course website:

<http://www.utstat.toronto.edu/brunner/oldclass/441s24>