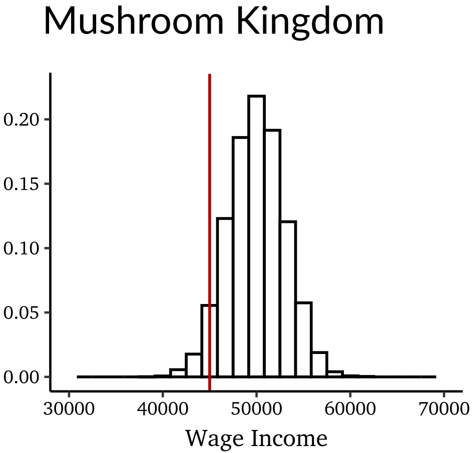
ECON 340
Economic Research Methods

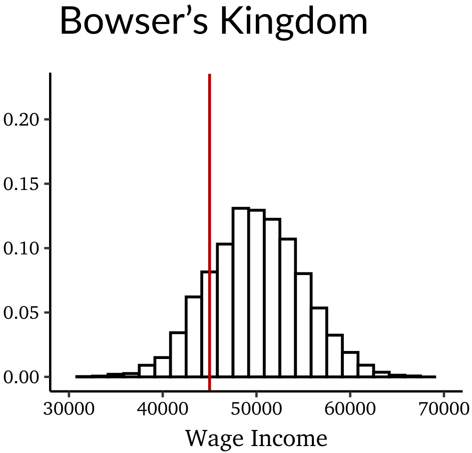
Div Bhagia

Lecture 4
Covariance and Correlation

Where would you want to live?



Mean = Median= $50,000
SD= $3,000



Mean = Median= $50,000
SD= $5,000

Z-Score

We can calculate the Z-Score to capture how many standard deviations (σ) away from the mean (μ) a specific observation is.

$$Z = \frac{X - \mu}{\sigma} \quad \rightarrow \quad X = \mu + Z.\sigma$$

Example: $\sigma_{MK} = 3000$, $\sigma_{BK} = 5000$

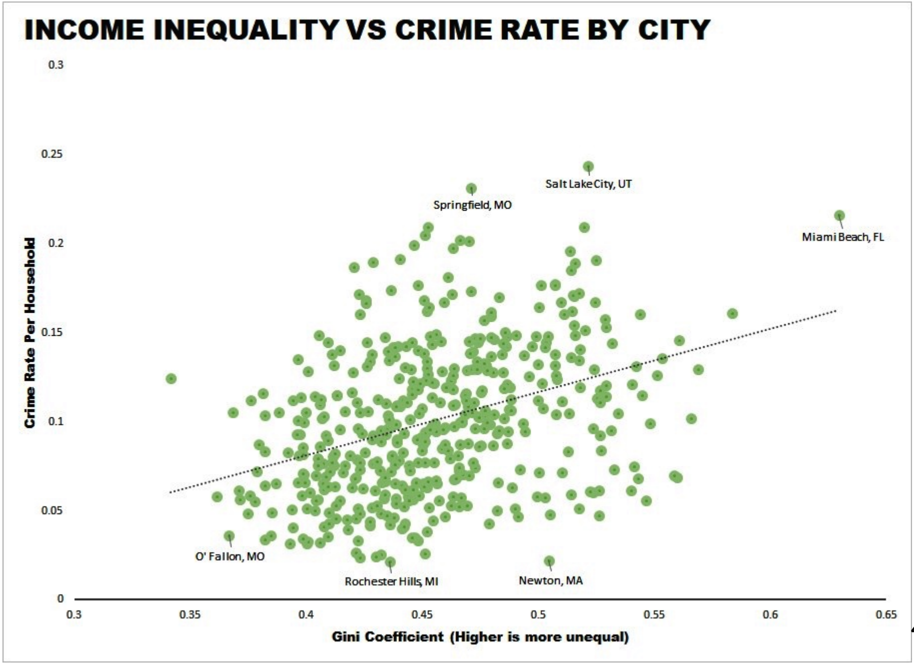
$$Z_{MK} = \frac{45000 - 50000}{3000} = -1.66 \qquad Z_{BK} = \frac{45000 - 50000}{5000} = -1$$

Describing Data

How do we summarize the information contained in a variable?

- Empirical distribution, histogram, percentiles
- Measures of central tendency: mean, median, mode
- Measures of variance: range, variance, standard deviation

How do we summarize the relationship between two variables?


INCOME INEQUALITY VS CRIME RATE BY CITY
Crime Rate Per Household
0.3
0.25
0.2
0.15
0.1
0.05
0
Salt Lake City, UT
Springfield, MO
Miami Beach, FL
O' Fallon, MO
Rochester Hills, MI
Newton, MA
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
Gini Coefficient (Higher is more unequal)

Describing Relationships

- Scatterplot: a graph where each point represents an observation of two variables
- Can see the relationship between two variables
- Positive relationship if when X is high Y is high (and when X is low Y is low)
- Negative relationship if when X is high Y is low (and when X is low Y is high)
- *How to construct a statistic to capture this?*

Covariance

Covariance indicates whether there is a positive or negative relationship between two variables.

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y) \quad (Population)$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \quad (Sample)$$

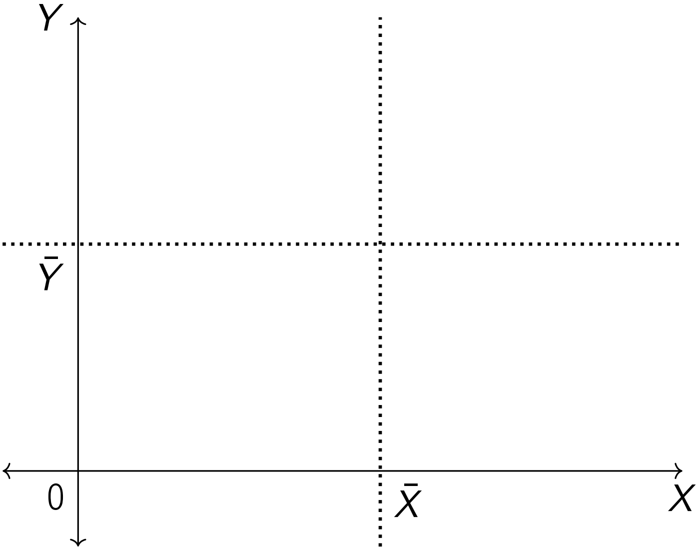
Example

X_i: sleep in hours, Y_i: exercise in hours

Week	X_i	Y_i	$(X_i - \mu_X)(Y_i - \mu_Y)$
1	6	0.5	
2	9	0.3	
3	9	1	
Total			

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y) =$$

Why does the formula work?



Correlation

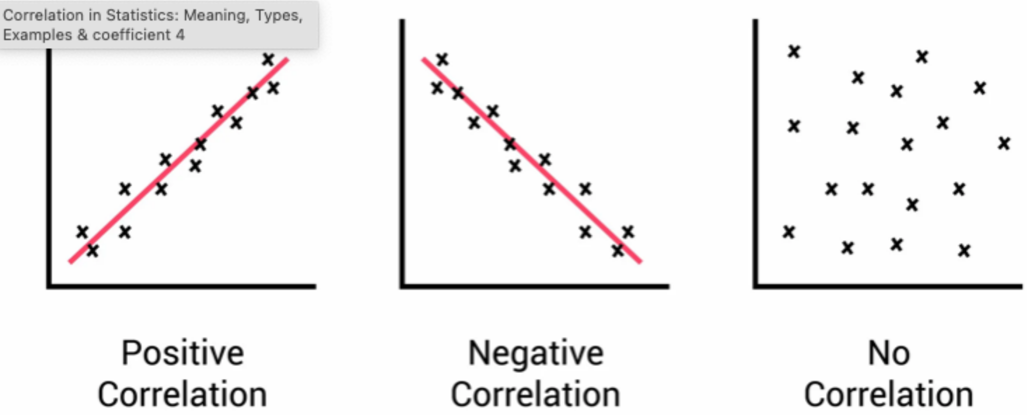
Correlation also indicates the *strength* of the relationship in addition to the *direction*.

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (Population) \qquad r_{XY} = \frac{S_{XY}}{S_X S_Y} \quad (Sample)$$

Bounded between -1 and 1.

- $\rho = 0$, no linear relationship
- $\rho = 1$, perfect positive linear relationship
- $\rho = -1$, perfect negative linear relationship

Correlation



Example

X_i: sleep in hours, Z_i: exercise in minutes

Week	X_i	Z_i	$(X_i - \mu_X)(Z_i - \mu_Z)$
1	6	30	
2	9	18	
3	9	60	
Total			

$$\sigma_{XZ} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Z_i - \mu_Z) =$$

Finally... Correlation is not causation

A positive correlation between inequality and crime doesn't suggest that inequality $\rightarrow$ crime. This is for two reasons:

- *Reverse causality*: crime $\rightarrow$ inequality (unlikely here but a concern in many situations)
- *Other confounding factors*: larger, more congested cities tend to be more unequal and also have higher crime rates

Things to do next

- Let me know the members of your research group by the end of the day
- Install R and R Studio on your computers if you haven't already (How to handout on Canvas)
- Please read Chapter 1 from Stock & Watson (uploaded on Canvas). We will discuss it on Tuesday.
- Coming up: Problem Set 1 (Due on Tues, 09/05)