INTR 5057 Research Design & Methods

Juraj Medzihorsky

C E U

Day 11, 2016-12-02

Homework #1

- Almost done grading. Overall OK.
- Keep perspective when you see your grade.
- Common issues:
  - Unfocused Qs.
  - Focus on single cases even for "effects of causes" Qs.
  - Mixing up Qs and Hs.
  - Causes vs. conditions.

Summarizing a Single Variable

How to summarize the following information?

x
1
4
0
3
3
2

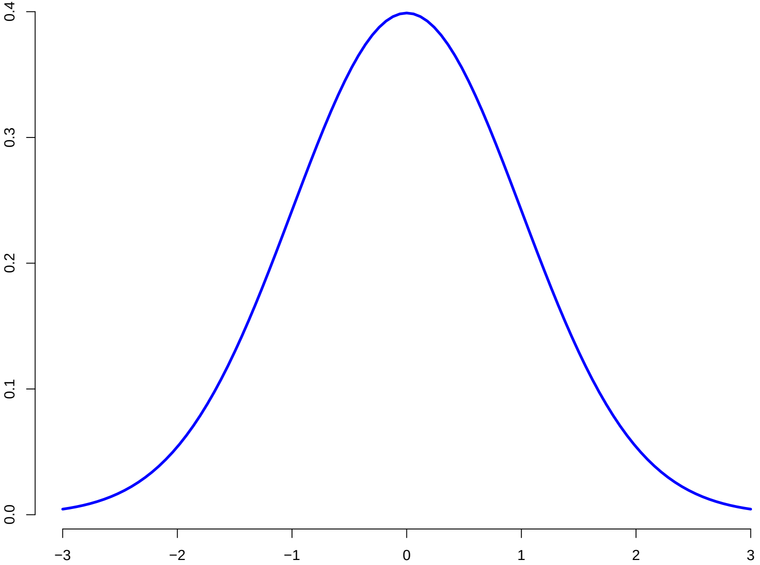
Summarizing a Single Variable

- Average (mean)
- Mode
- Median
- Midrange

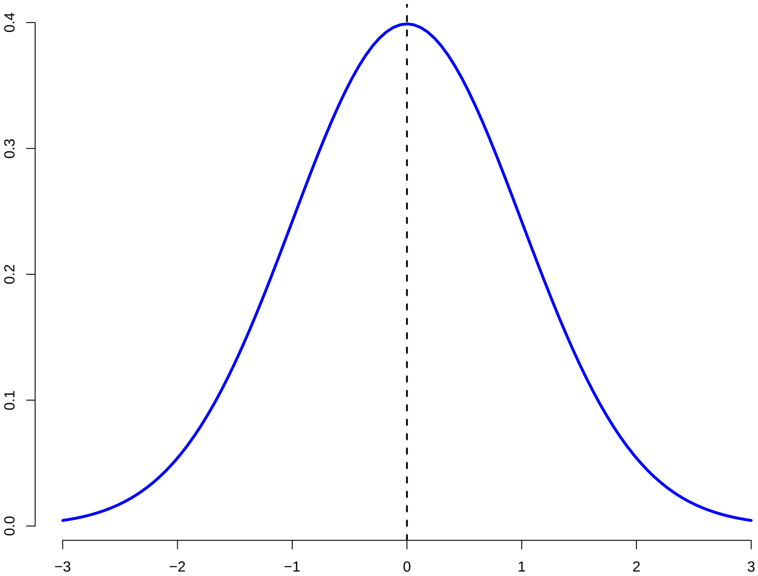
Summarizing a Single Variable

- Average: Sum divided by the number of values.
- Mode: The most common value.
- Median: Half of the observations have less, half more.
- Midrange: Midpoint between maximum and minimum.

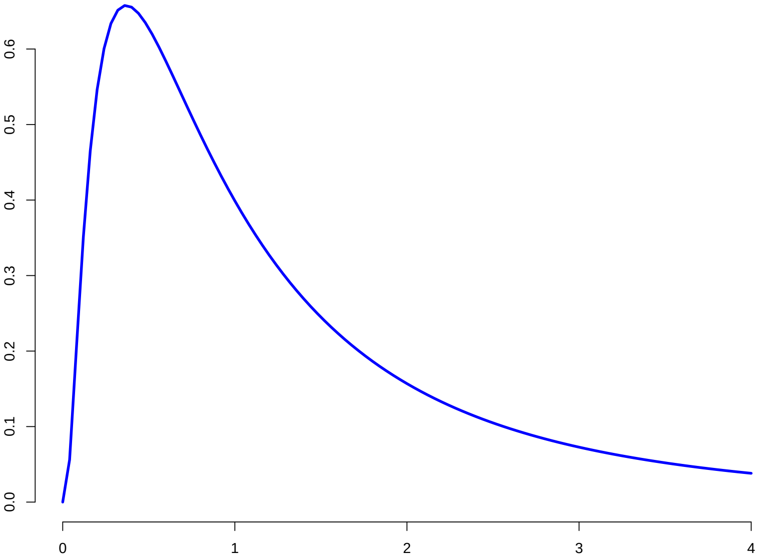
A Continuous Variable



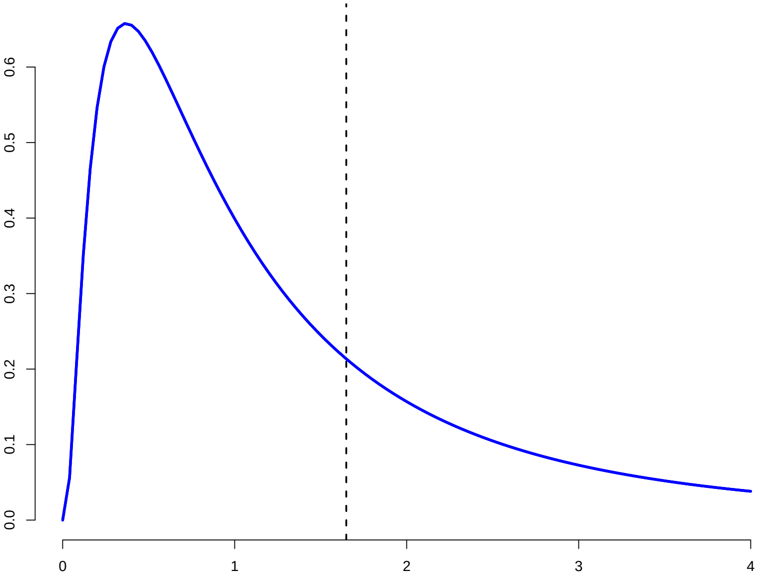
Mean, Median, Mode, Midrange



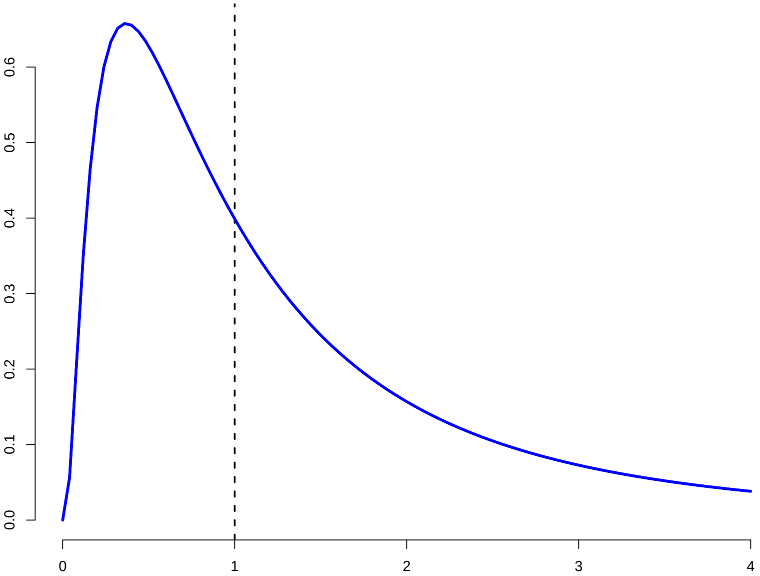
A Continuous Variable



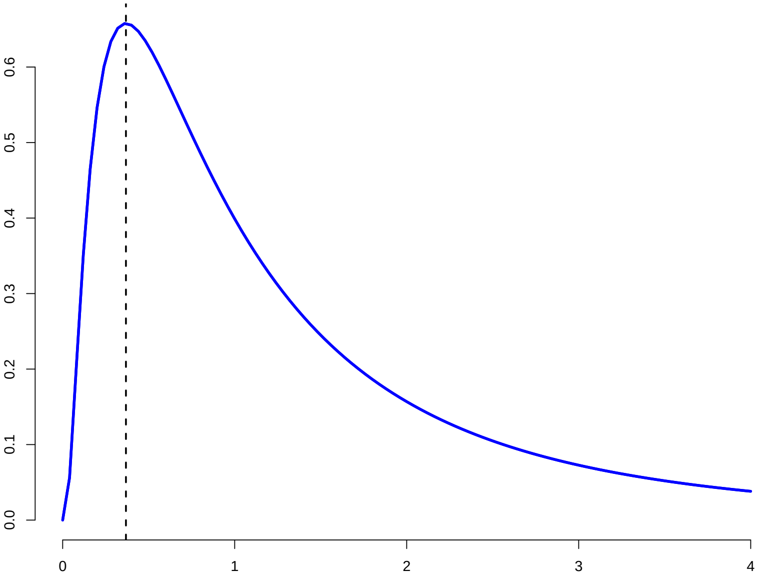
Mean



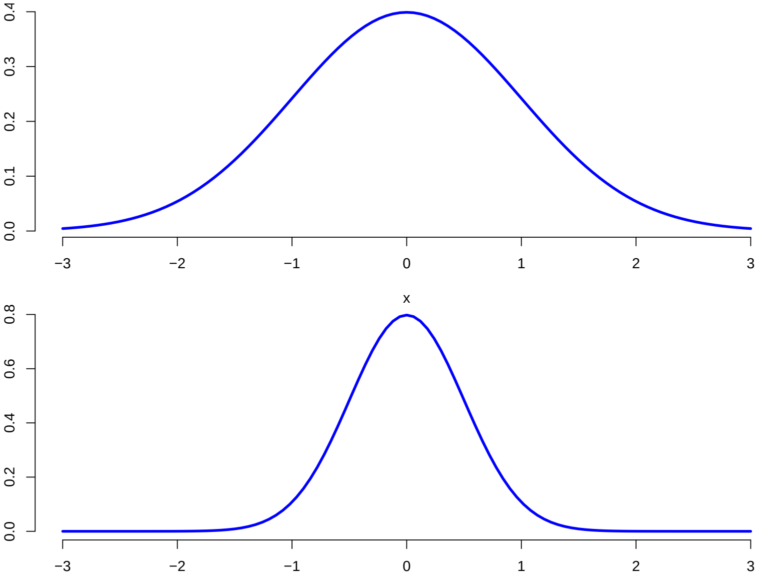
Median



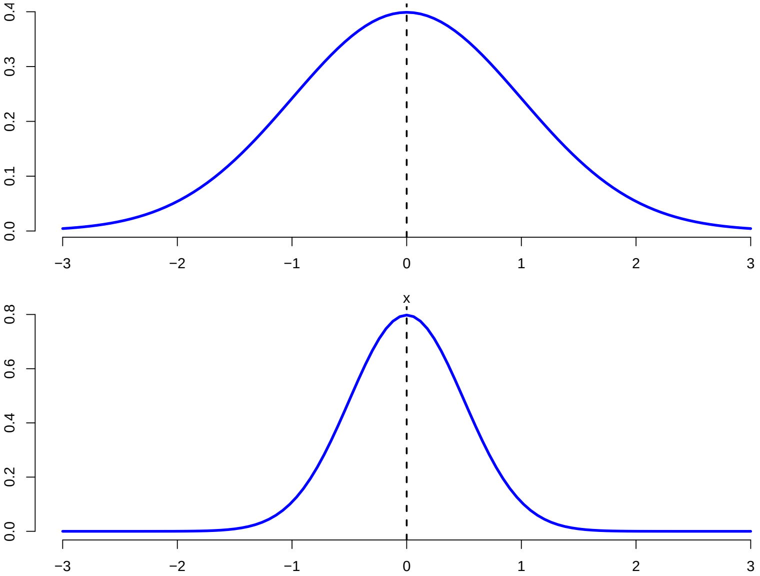
Mode



Two Continuous Variables



Same Mean



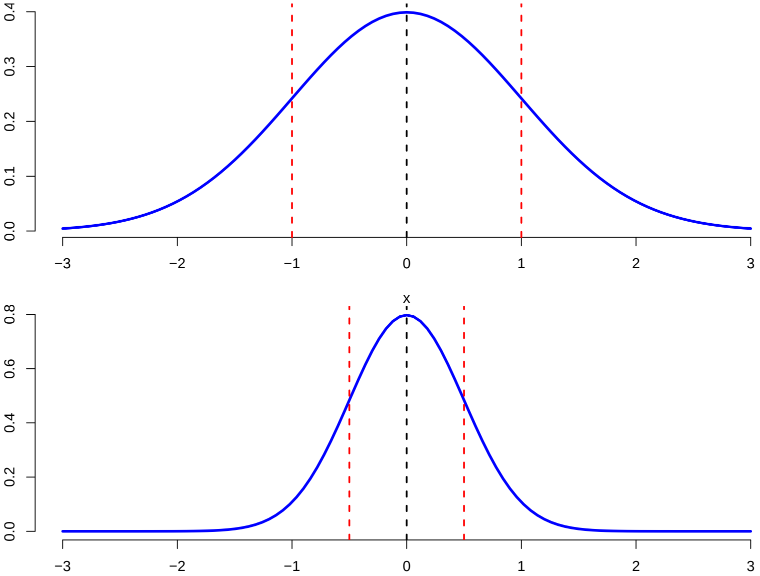
Standard Deviation

A measure of dispersion from the average.

Square root of the average squared distance from the average.

$$\sigma = \sqrt{\frac{\sum(\mu - x_i)^2}{N}}$$

Standard Deviation



Summarizing Two Variables

How to summarize the following information?

x	y
1	1
0	1
1	0
1	0
0	0
1	0

Summarizing Two Variables

- Summarize each of them separately.
- Capture information about their **association.**

Cross Table

		y	
		0	1
x	1	3	1
	0	1	1

Cross Table

		y	
		0	1
x	1	a	b
	0	c	d

Cross Product Ratio

A measure of association between two binary variables also know as odds ratio.

If cross product ratio $= 1$ then the variables are independent.

$$cpr = \frac{a \times d}{b \times c}$$

Cross Sum Ratio

An alternative to the cross product ratio.

$$csr = \frac{a + d}{b + c}$$

Relative Risk Ratio

A measure of association between two binary variables.

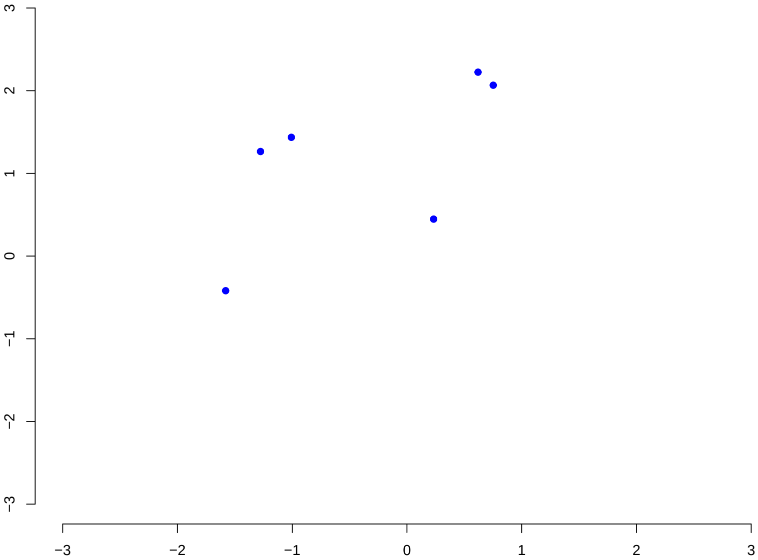
$$rr = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Summarizing Two Variables

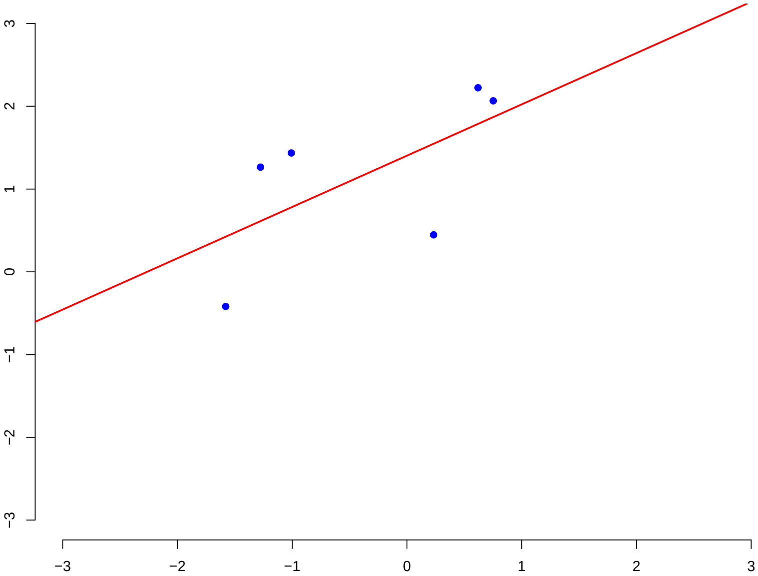
How to summarize the following information?

y	x
0.8	2.1
0.6	2.2
-1.6	-0.4
0.2	0.5
-1.3	1.3
-1.0	1.4

Scatter Plot



Correlation



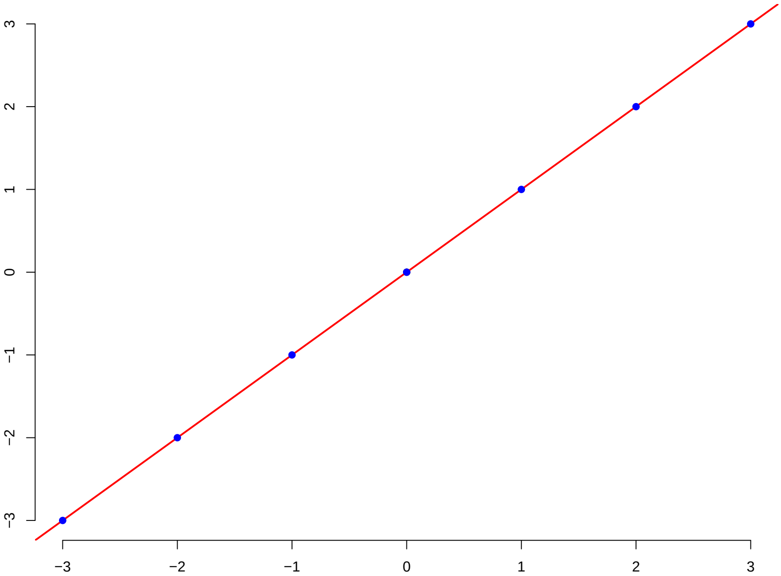
Correlation

A measure of association between two continuous variables.

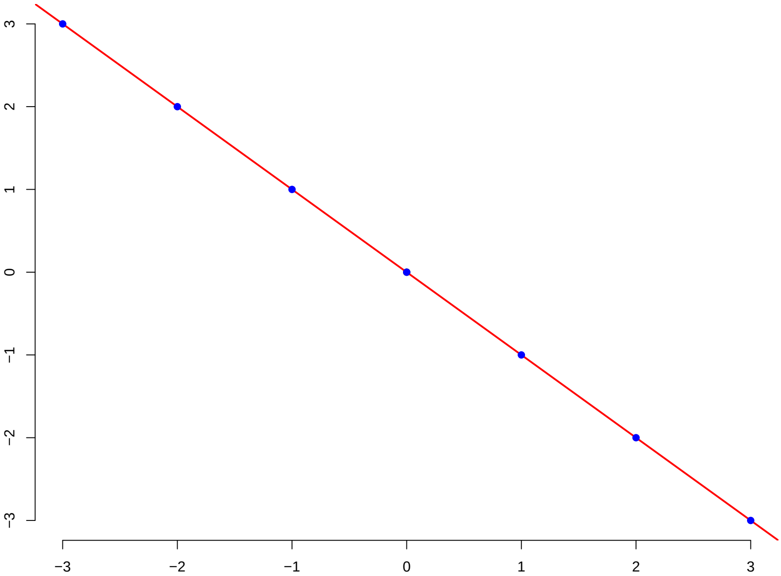
$$\rho_{x,y} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

Ranges from -1 to 1 on a closed interval.

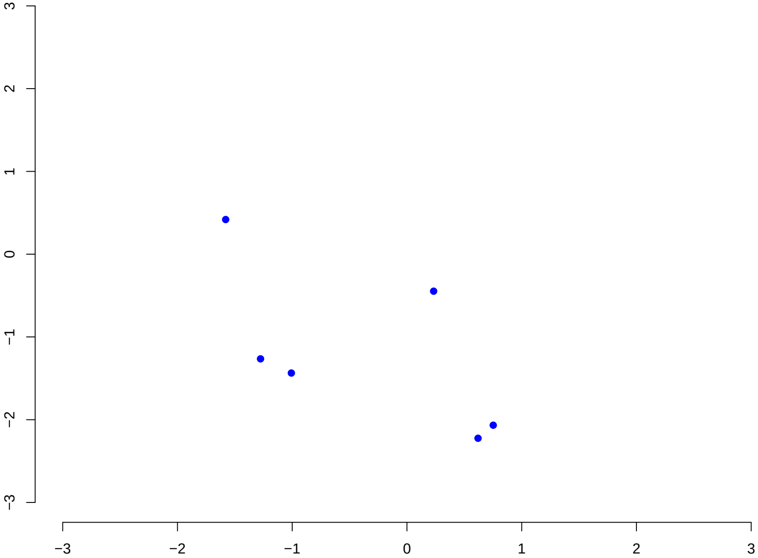
$\rho = 1$



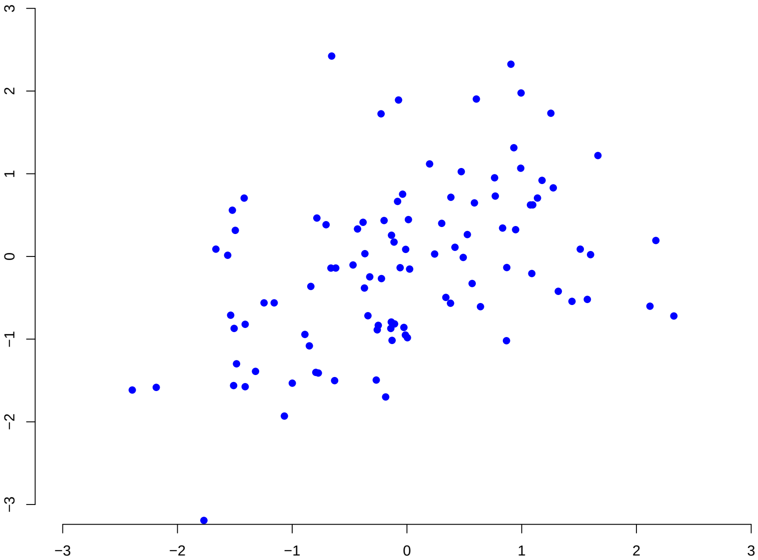
$$\rho = -1$$



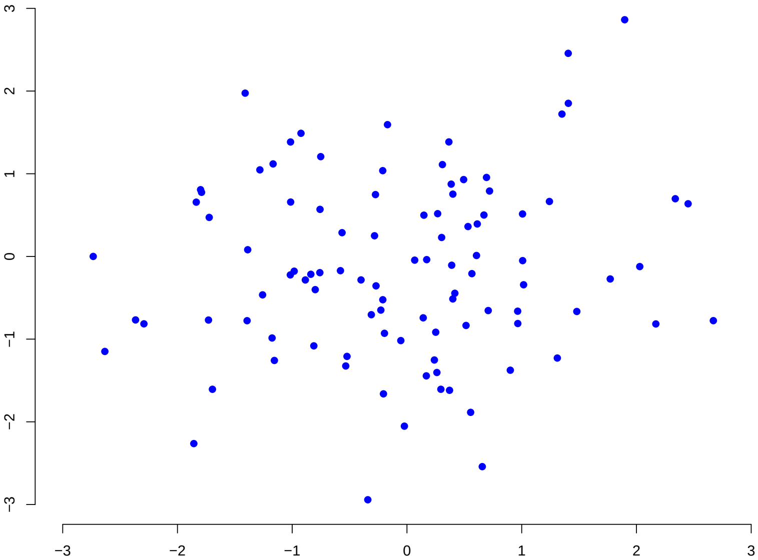
$$\rho = -0.63$$



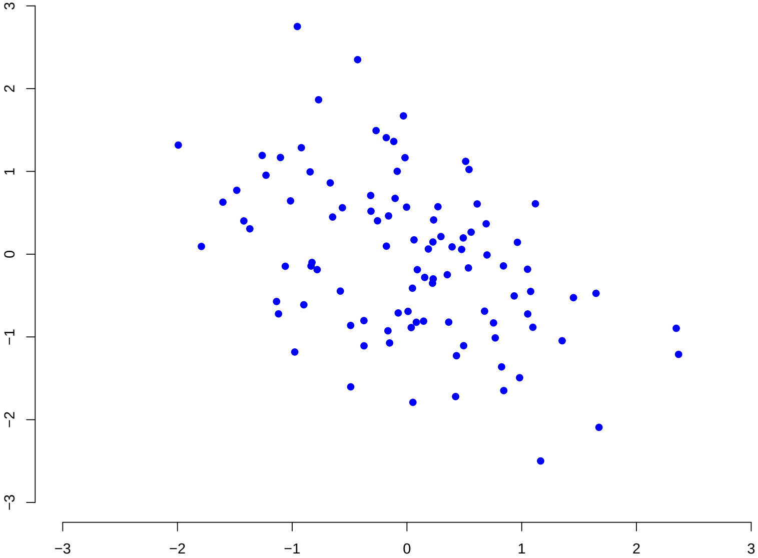
$$\rho = 0.44$$



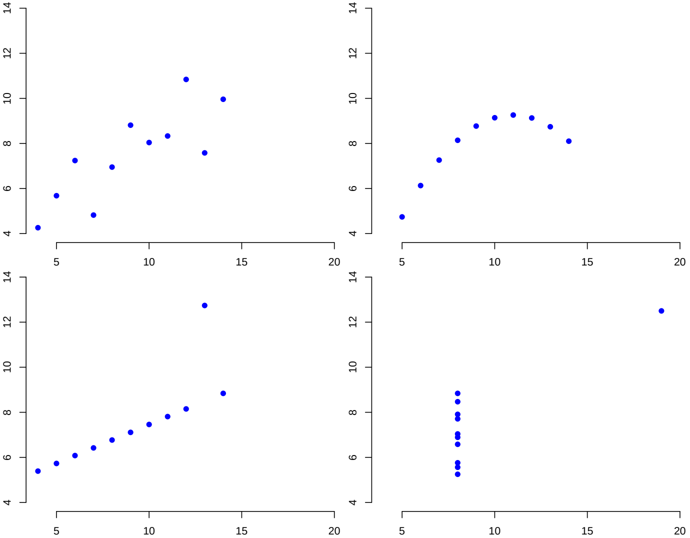
$$\rho = 0.09$$



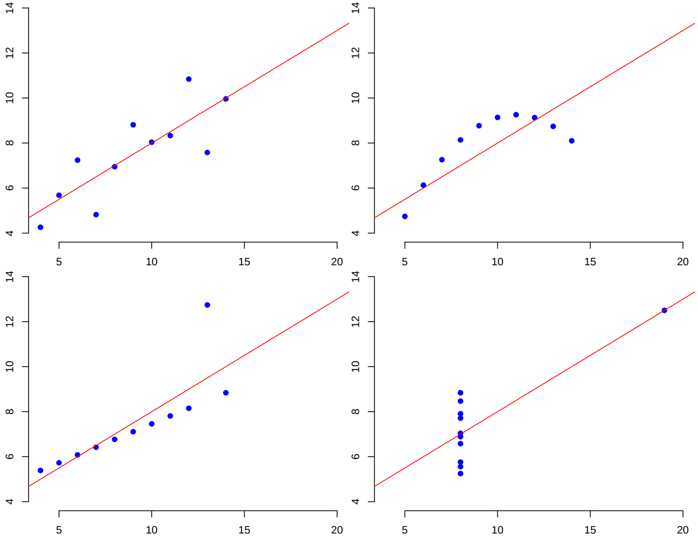
$$\rho = -0.46$$



Anscombe's Quartet



$$\rho = 0.82$$



Simpson's Paradox

The whole sample:

	heal	didn't
drug	20	20
no drug	16	24

Simpson's Paradox

Females:

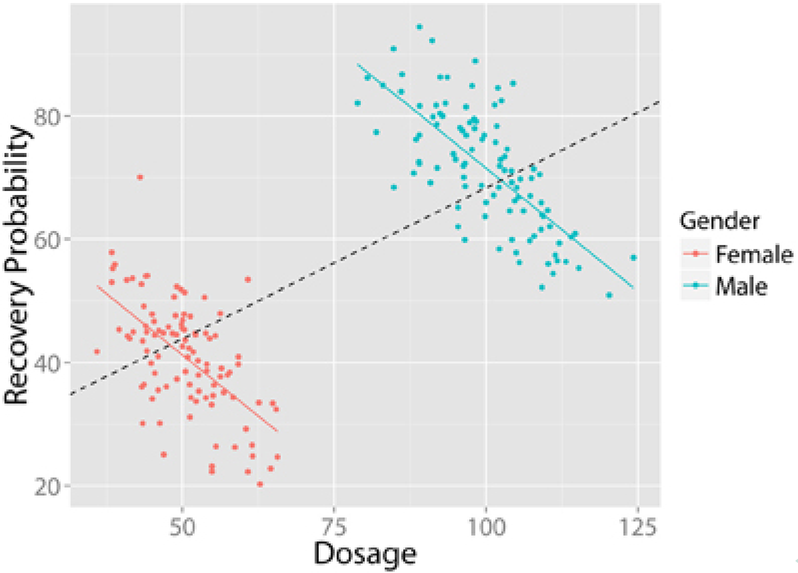
	heal	didn't
drug	2	8
no drug	9	21

Simpson's Paradox

Males:

	heal	didn't
drug	18	12
no drug	7	3

Simpson's Paradox



Simpson's Paradox

- In the whole population association in one direction.
- In subsets of the population association in the opposite direction.
- Not really a paradox when you think about it.
- A serious problem is that people rush ahead with causal interpretations.

Association & Causality

- Non-statisticians say *"correlation does not imply causation."*
- Statisticians say *"association does not imply causation."*
- Calling all association "correlation" is like calling all motor vehicles "cars."

Goals

Goals

- Describe.
- Explain.
- Predict/Forecast.
- ...

Table 1: Electoral and Replacement Volatility in Post-Communist Europe

	Verification		Without Bosnia-Herzegovina		Corrected Bosnia-Herzegovina	
	Electoral Volatility	Replacement Volatility	Electoral Volatility	Replacement Volatility	Electoral Volatility	Replacement Volatility
GDP Change from 1989	0.639	−4.623***	0.116	−6.066	0.004	−6.002
	(0.693)	(1.326)	(3.206)	(7.178)	(3.233)	(6.609)
GDP Change Between Elections	−2.059	9.019	−1.891	6.677	−1.076	4.576
	(5.219)	(10.128)	(5.898)	(10.704)	(5.229)	(10.064)
Effective Number of Electoral Parties	0.446	−0.346	0.452	−0.264	0.471	−0.462
	(0.313)	(0.533)	(0.316)	(0.558)	(0.326)	(0.546)
Log Weighted District Magnitude	−0.784	0.638	−0.789	0.603	−0.824	0.820
	(0.887)	(2.931)	(0.882)	(2.893)	(0.886)	(2.872)
Presidential System	−4.631	6.784	−4.847	5.532	−4.928	6.659
	(4.126)	(9.435)	(4.606)	(10.241)	(4.623)	(10.296)
Semi-Presidential System	−2.788	4.255	−2.813	4.017	−2.596	2.621
	(2.211)	(5.897)	(2.286)	(5.885)	(2.266)	(5.887)
Proportional Representation	0.827	0.077	0.852	−0.146	0.987	−0.739
	(2.228)	(6.004)	(2.265)	(5.943)	(2.223)	(5.948)
Ethnic Fractionalization	−6.163	−2.677	−6.716	−5.298	−5.713	−11.828
	(6.397)	(18.978)	(6.784)	(22.939)	(6.772)	(22.931)
Years Since Collapse of Communism	0.848	−2.633	0.828	−1.989	0.732	−1.959
	(0.807)	(2.153)	(0.863)	(2.117)	(0.797)	(1.976)
Years Since Collapse Squared	−0.031	0.070	−0.029	0.045	−0.026	0.049
	(0.042)	(0.101)	(0.044)	(0.097)	(0.043)	(0.093)
Constant	13.059**	41.941***	13.586***	43.661***	12.885**	48.191***
	(5.318)	(13.329)	(5.115)	(14.057)	(5.034)	(14.509)
Countries	21	21	20	20	21	21
Pairs of Elections	89	89	86	86	89	89
R^2	0.116	0.139	0.114	0.119	0.112	0.109

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$ (two-tailed).

Regression

- Whether we like it or not, **regression** is the workhorse of quantitative social science.

Regression

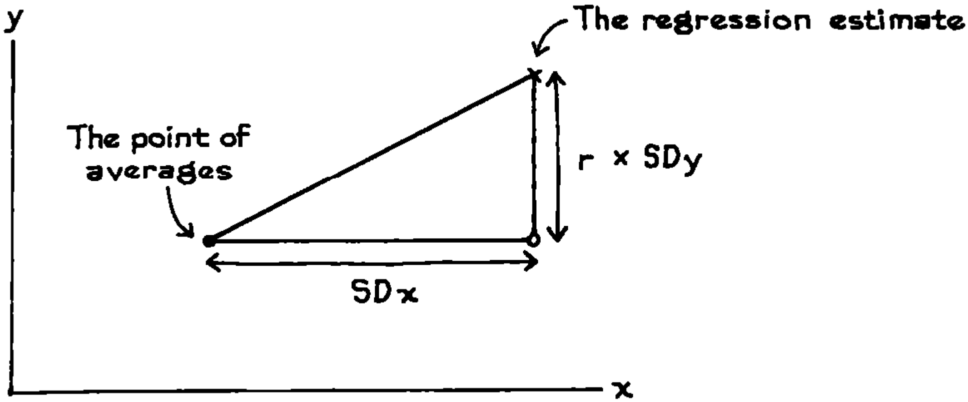
- Whether we like it or not, **regression** is the workhorse of quantitative social science.
- Any previous experiences with regression?

Regression

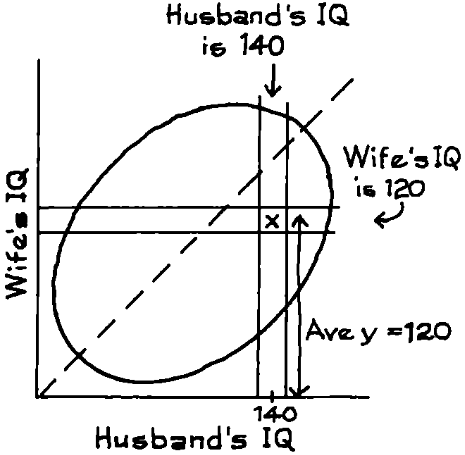
- One variable as a function of one or more other variables.
- **Conditional association**.
- Typically used to **explain** or **predict**.

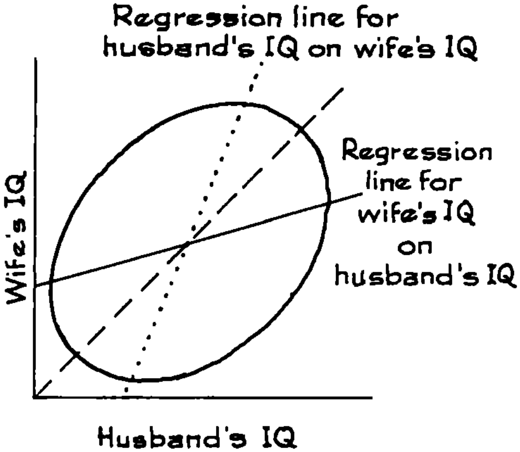
Regression

Figure 2. Regression method. When x goes up by one SD, the average value of y only goes up by r SDs.



Regression


Husband's IQ
is 140
Wife's IQ
is 120
x
Ave y =120
Wife's IQ
140
Husband's IQ


Regression line for
husband's IQ on wife's IQ
Regression
line for
wife's IQ
on
husband's IQ
Wife's IQ
Husband's IQ

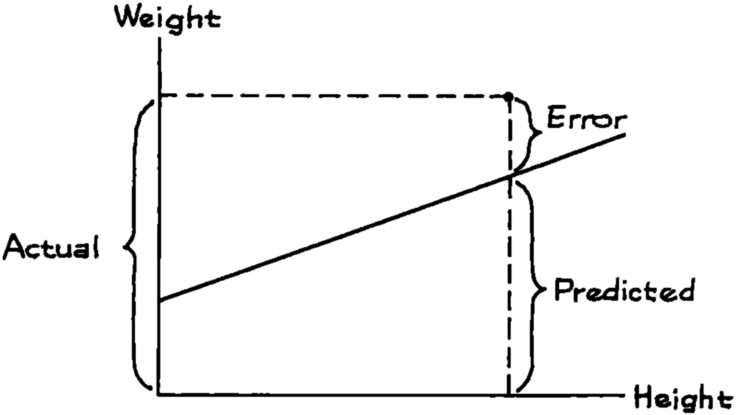
Linear Regression

$$y_i = \alpha + \beta \times x_i + \epsilon_i$$

- y: LHS, "dependent variable," outcome
- x: RHS, "indepdent variable," predictor, determinant
- α: intercept, "constant"
- β: slope, coefficient, "effect"
- ϵ: residual, "error"

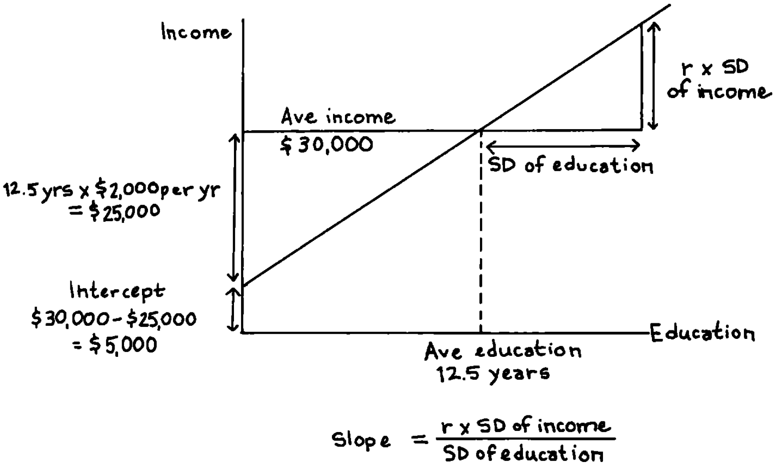
Regression

Figure 2. Prediction error equals vertical distance from the line.



Regression

Figure 3. Finding the slope and intercept of the regression line.



Linear Regression

$$y_i = \alpha + \beta x_i + \epsilon_i$$

- Ordinary Least Squares (OLS): $\Sigma_i \epsilon_i^2$.
- Probabilistic I.:

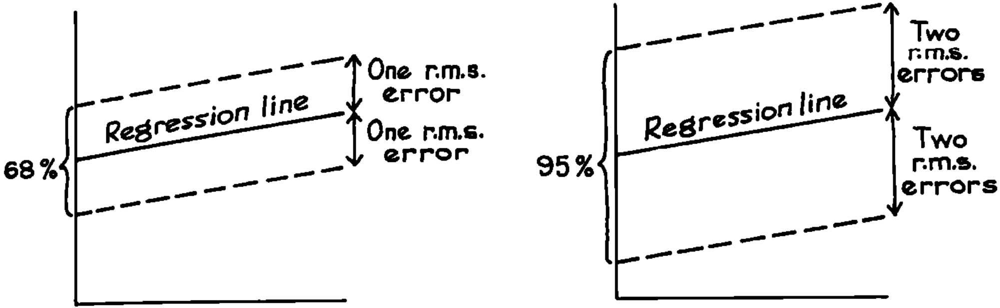
$$\epsilon_i \sim Normal(0, \sigma)$$

- Probabilistic II.:

$$y_i \sim Normal(\alpha + \beta x_i, \sigma)$$

Regression

Figure 3. Rule of thumb. About 68% of the points on a scatter diagram fall inside the strip whose edges are parallel to the regression line, and one r.m.s. error away (up or down). About 95% of the points are in the wider strip whose edges are parallel to the regression line, and twice the r.m.s. error away.



Regression

Figure 10. A football-shaped scatter diagram. Take the points inside a narrow vertical strip. Their y-values are a new data set. The new average is given by the regression method. The new SD is given by the r.m.s. error of the regression line. Inside the strip, a typical y-value is around the new average—give or take the new SD.