

## Causality and Lack Thereof

1. A clear **temporal precedence** of the cause before the effect is necessary for a causal relationship:

$$effect_t \leftarrow cause_{t-1}$$

2. The role of the dependent variable and the independent variable may **switch** depending on the **perspective**. E.g., the parent's income has an effect on the children's education, but also a person's education level influences her/his income potential:

$$income \leftarrow education$$

or

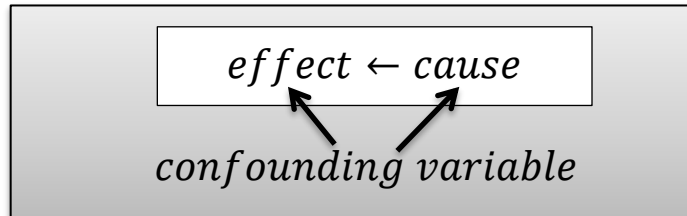
$$education \leftarrow income$$

Note: the parent's income precedes the child's education and a person's education precedes a her/his income potential.

3. Correlation may be caused by plain **random coincidences**. Statistical significance tests will help quantifying the potential impact of these random effects.

$$effect \xleftarrow{random} cause$$

4. A **lurking third variable**, jointly related with the two variables under investigation, may induce a relationship. This is known as **spurious** correlation or **confounding** effect.



## Covariation Relationships

- This lecture discusses bivariate **covariation** between pairs of variables. In contrast, regression analysis assumes a cause-effect relationship.
- The focus of bivariate analysis is to explore how values taken by one variable **co-vary** with values taken by another variable.
- Relationships are **not directed**, thus, one cannot say  $X_1$  influences  $X_2$ , i.e.,  $X_1 \rightarrow X_2$  or *vice versa*  $X_1 \leftarrow X_2$ . Both variables just mutually co-vary, i.e.,  $X_1 \leftrightarrow X_2$
- These bivariate co-variations can be explored
  - for nominal and ordinal scaled variables by nested bar charts (recall the spinogram) and quantified in contingency tables

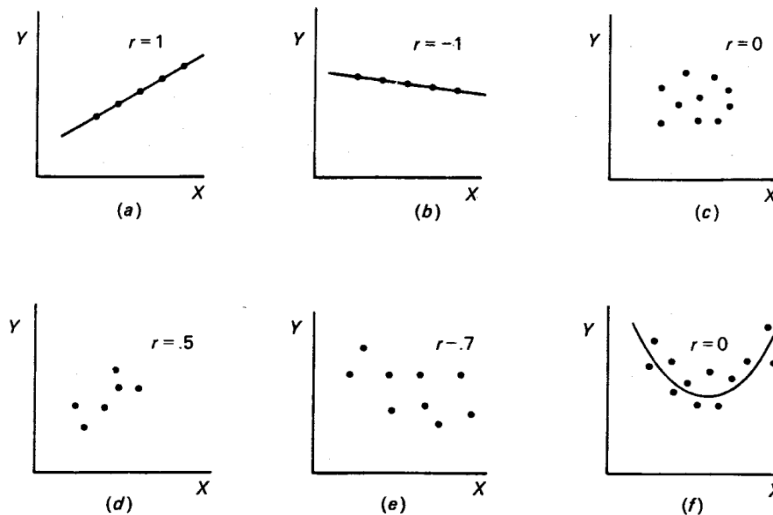
- for quantitative variables we can scatterplot both variables and quantify their linear relationship with the **Pearson correlation coefficient** and non-linear but monotonic relationships by the **Spearman rank correlation coefficient**.
- For **mixed** qualitative and quantitative variables the co-variation of a quantitative variable can be investigated by **stratifying its conditional** statistics and distributions by the **categories** of the qualitative variable (recall the mean bar plot and parallel box-plots).
- Def. Statistical Dependence: When the probability of a variable taking a particular value is influenced by a given value of another variable, then these two variables are statistically dependent.

## Correlation Analysis

- Correlation between two metric variables measures their **linear** relationship:
  - *Positive* linear relationship means that as variable  $X_1$  increases so does the variable  $X_2$ .
  - *Negative* linear relationship means that as  $X_1$  increases variable  $X_2$  does decrease.
  - *Zero correlation* means that there is no systematic **linear** relationship between both variables.
  - This bivariate relationship must hold over the **full support** of  $X_1$  and  $X_2$ , i.e., it cannot be piecewise for specific subsets of intervals.

Unless, however, a qualitative variable allows **stratifying** the relationship down into subgroups.

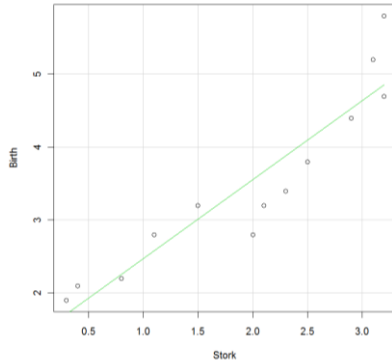
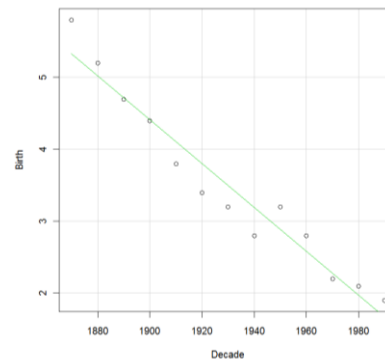
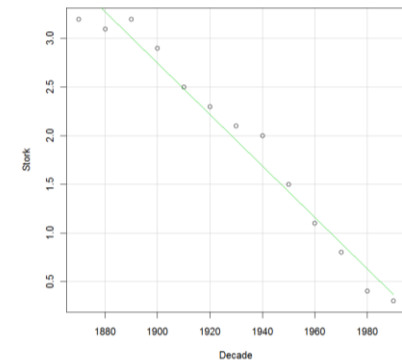
- The relationship between two metric variables is best displayed in a scatterplot. This is also an important **exploratory tool**



## Meaningfulness of the Correlation Coefficient

- A statistical relationship *per se* does not need to be meaningful.
  - Example: **Confounding variable**.  
Birth rates and the density of storks in a rural German province over several decades in the last century:

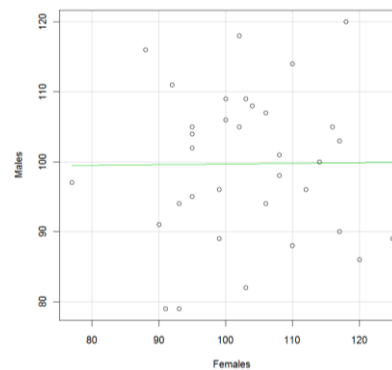
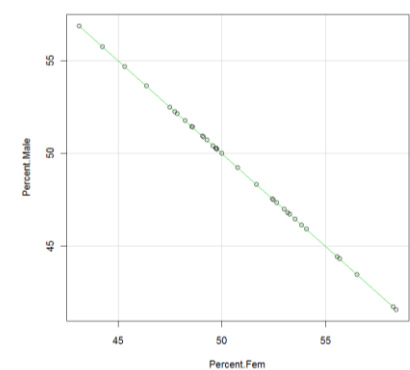


*Birth Rate by Stork Density**Birth Rate by Decade**Stork Density by Decade*

- Both the stork density and the birth rate vary by time. With time industrialization increases, this reduces the marshy habitat of storks and changes the reproductive behavior of the population.

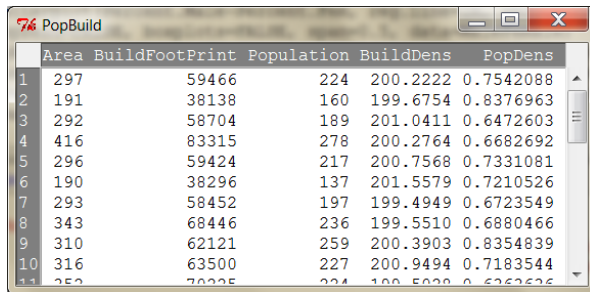
- Example: ***Induced Correlation***

	Males	Females	TotalPopulation	Percent.Fem	Percent.Male
1	97	77	174	44.25287	55.74713
2	95	95	190	50.00000	50.00000
3	90	117	207	56.52174	43.47826
4	86	120	206	58.25243	41.74757
5	102	95	197	48.22335	51.77665
6	91	90	181	49.72376	50.27624
7	109	103	212	48.58491	51.41509
8	82	103	185	55.67568	44.32432
9	116	88	204	43.13725	56.86275
10	114	110	224	49.10714	50.89286
11	96	112	208	53.84615	46.15385

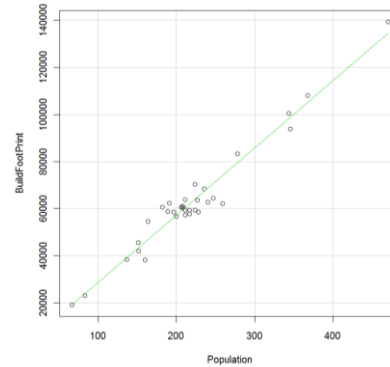
*Absolute Populations by Sex**%Males by %Females*

- Example: **Spurious correlation**

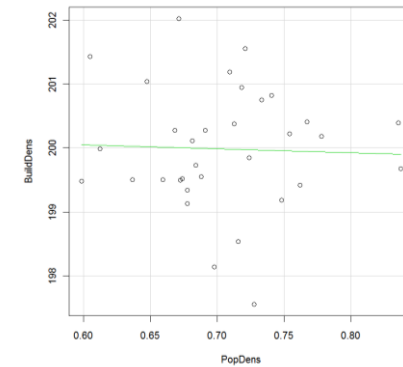
The variable "Size of Area" induces correlation. **Normalization** by the size of the area corrects for the absolute size effects.



	Area	BuildFootPrint	Population	BuildDens	PopDens
1	297	59466	224	200.2222	0.7542088
2	191	38138	160	199.6754	0.8376963
3	292	58704	189	201.0411	0.6472603
4	416	83315	278	200.2764	0.6682692
5	296	59424	217	200.7568	0.7331081
6	190	38296	137	201.5579	0.7210526
7	293	58452	197	199.4949	0.6723549
8	343	68446	236	199.5510	0.6880466
9	310	62121	259	200.3903	0.8354839
10	316	63500	227	200.9494	0.7183544



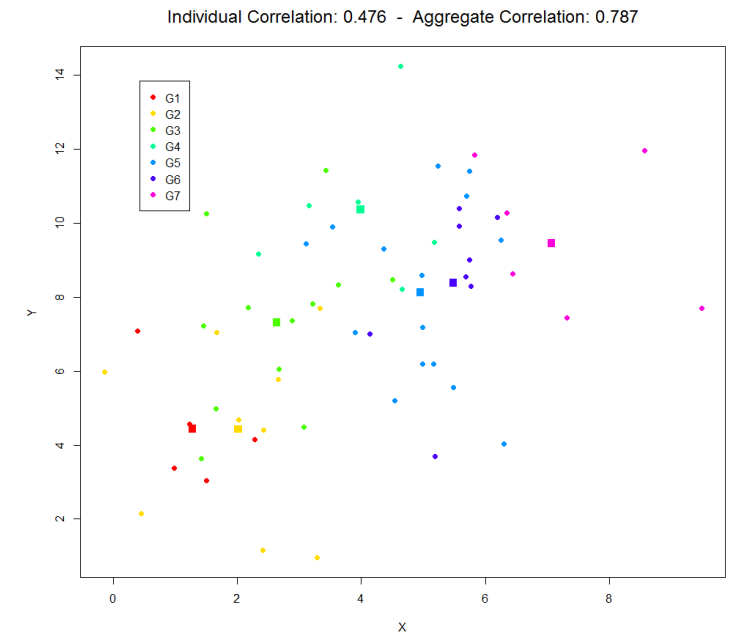
Absolute numbers



Relative Density Numbers

- Example: **Aggregation effects**

Correlation of individual observations (e.g., individuals living within regions) visualized by dots versus correlation of aggregated observations (e.g., averaged individual data in regions) visualized by squares. => This is known as the modifiable areal unit problem (MAUP)



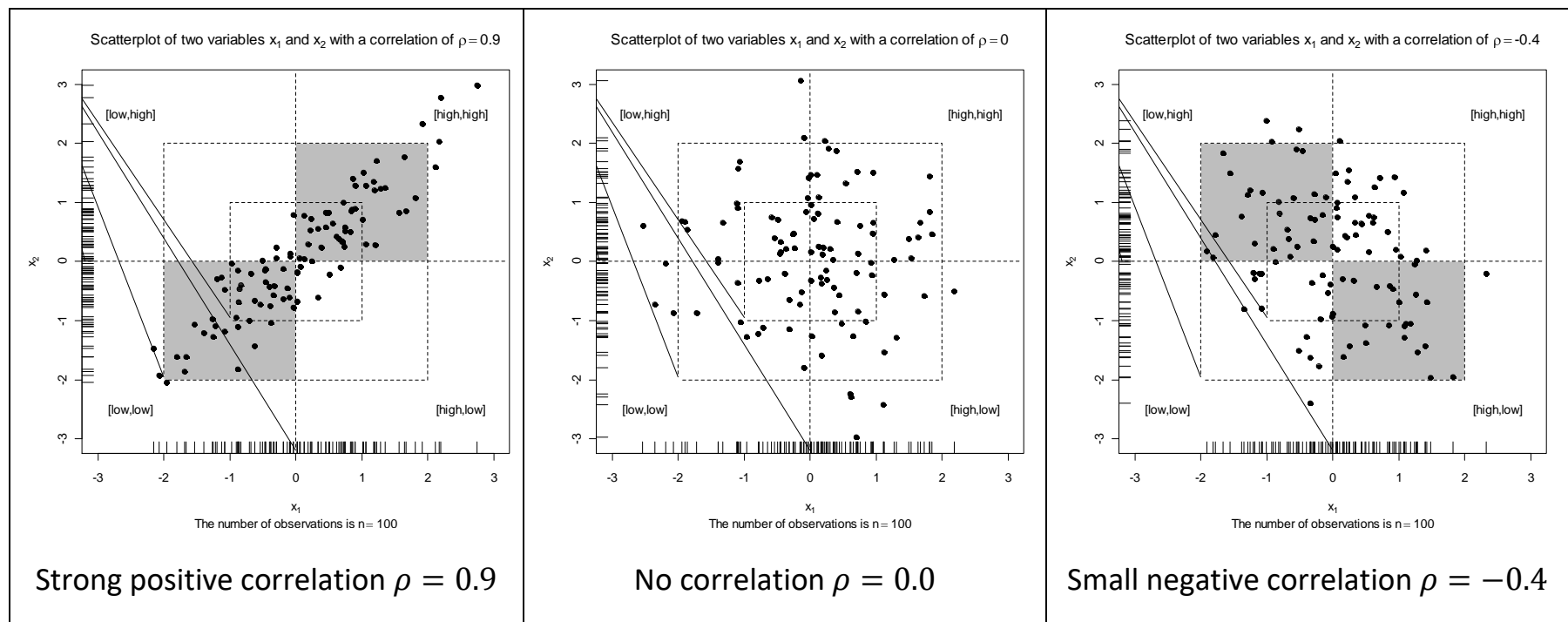
## Structure of Pearson's Product-Moment Correlation Coefficient

- Empirically the joint variation between two variables is expressed by their measure of

**covariance:**  $s_{x_1, x_2} = \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1) \cdot (x_{i2} - \bar{x}_2)}{n - 1}$

Note that the denominator remains  $(n - 1)$ .

- Geometric interpretation: Point pairs  $[x_1, x_2]$  in quadrant defined by their variation  $[(x_{i1} - \bar{x}_1), (x_{i2} - \bar{x}_2)]$  around their means  $\bar{x}_1$  and  $\bar{x}_2$



- The covariance is **not standardized** to a comparable value range. It, therefore, it becomes difficult to compare for different pairs of variables.

This is because the scale of its parent variables  $X_1$  and  $X_2$  may differ.

- For these reasons the covariance is **standardized** by the standard deviations  $s_{x_1}$  and  $s_{x_2}$  of its parent variables.

This gives the **correlation** coefficient: 
$$r = \frac{s_{x_1, x_2}}{s_{x_1} \cdot s_{x_2}} = \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1) \cdot (x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 \cdot \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2}}$$

- Notes:
  - The value range of the correlation coefficient is restricted within the interval  $r \in [-1, 1]$ . Negative values imply a negative **linear** relationship, a value of zero implies no relationship and positive value implies a positive **linear** relationship.
  - Then denominator  $n - 1$  in the covariance and the standard deviations cancels out.
  - BBR give an equivalent computational equation on p 168.

## The Spearman's Rank Correlation Coefficients

- The distribution of a variable does not need to be symmetric or without outliers. **Ranks are not affected by skewness or outliers.**

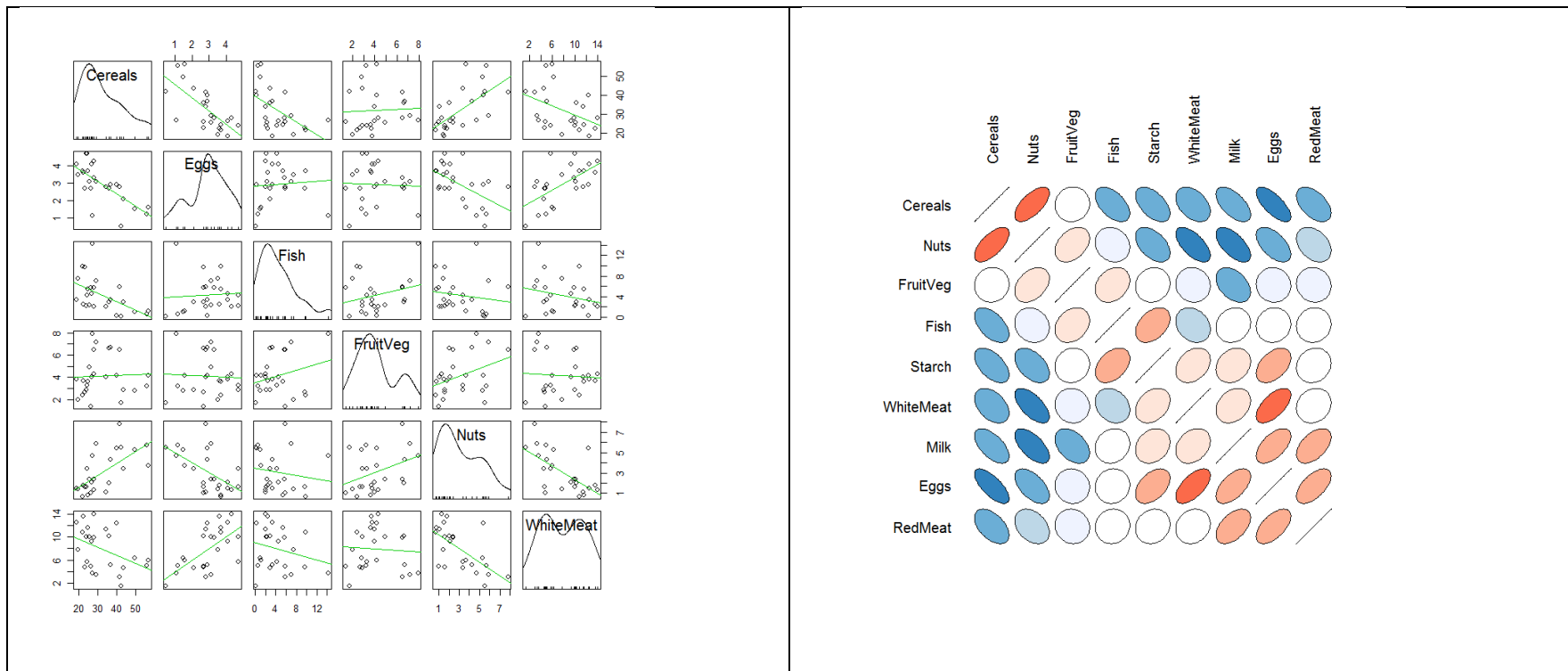


- **Replacing** the observations in each variable by their associated ranks and then calculating the correlation between these ranks leads to the Spearman's rank correlation coefficient.
- Spearman's rank correlation coefficient is **robust** against skewness and outliers.
- It measures **monotonically** increasing or decreasing relationships, which do not necessarily need to follow a straight line. However, it cannot measure complex non-linear relationships properly.

## Scatterplot Matrix

There is *a priori* no reason to believe that the protein consumption of different food groups in 25 European countries before the fall of the iron curtain exhibits directional influence.

	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	FruitVeg	Region	TotalProtein
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7	Balkan	71.2
Austria	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3	WestEuro	86.4
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0	WestEuro	87.3
Bulgaria	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2	Balkan	90.6
Czechoslovakia	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0	EastEuro	82.8
Denmark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4	Scania	89.8
E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6	EastEuro	75.7
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4	Scania	90.4
France	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5	WestEuro	98.2
Greece	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5	Mediterra	97.7
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2	EastEuro	84.3
Ireland	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9	WestEuro	91.3
Italy	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7	Mediterra	84.0
Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7	WestEuro	84.7
Norway	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7	Scania	81.7
Poland	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6	EastEuro	92.7
Portugal	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9	Iberia	75.6
Romania	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8	Balkan	86.9
Spain	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2	Iberia	77.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0	Scania	80.0
Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9	WestEuro	88.1
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3	WestEuro	88.4
USSR	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9	EastEuro	91.9
W Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8	WestEuro	79.3
Yugoslavia	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2	Balkan	88.5



## Spurious Relationships in Cross-Tabulations

Source: Kaplan, Daniel T, 2009. Statistical Modeling. A Fresh Approach. Self-published. p 151f

According to a nautical fiction, a research assistant was asked to explore for the U.S. Coast Guard whether life vests are saving lives?

The research assistant sampled 500 records of man-over-board accidents on large vessels with the following result:

Rescue \* Vest Crosstabulation

			Vest		Total
			Wearing	No Vest	
Rescue	Survived	Count	100	71	171
		% within Vest	25.0%	71.0%	34.2%
	Drowned	Count	300	29	329
		% within Vest	75.0%	29.0%	65.8%
Total		Count	400	100	500
		% within Vest	100.0%	100.0%	100.0%

Conclusions: Apparently wearing a vest is ***increasing the likelihood of drowning***.

The supervisor felt extremely uncomfortable with these results and asked the research assistant to investigate if weather conditions may have an effect on the survival chances.

The partial results broken down by the weather conditions tell a different story, more in tune with common sense:

Rescue \* Vest \* Weather Crosstabulation

Weather				Vest		Total
				Wearing	No Vest	
Fair	Rescue	Survived	Count	19	70	89
			% within Vest	95.0%	87.5%	89.0%
	Drowned	Count		1	10	11
			% within Vest	5.0%	12.5%	11.0%
	Total	Count		20	80	100
			% within Vest	100.0%	100.0%	100.0%
Foul	Rescue	Survived	Count	81	1	82
			% within Vest	21.3%	5.0%	20.5%
	Drowned	Count		299	19	318
			% within Vest	78.7%	95.0%	79.5%
	Total	Count		380	20	400
			% within Vest	100.0%	100.0%	100.0%

Conclusions: Under both weather conditions wearing a vest increases the likelihood of survival.

Explanation: In foul weather, sailors have the habit of putting life vests on, whereas in fair weather they mostly do not bother doing so. However, falling over board in foul weather diminishes the likelihood of survival irrespectively of wearing a vest or not (see red numbers).

Apparently, the weather conditions are correlated with the status of wearing a vest and, therefore, the survival likelihood.