

*JMP Discovery Summit Europe 2023*

# Revisiting Pearl's Influenza Studies Using JMP®

**Roselinde Kessels**, Assistant Professor, Maastricht  
University

**Chris Gotwalt**, Chief Data Scientist, JMP

**Guido Erreygers**, Professor of Economics, University of  
Antwerp

---

TOPIC: DATA EXPLORATION

LEVEL: 

---

# Overview

- The “Spanish Flu” pandemic
- Pearl’s influenza studies I and II
- Pearl’s data and Census data
- Variable selection with null factor and bootstrap simulation
- Discussion of results
- Conclusion



*Raymond Pearl.*  
1879-1940

# The Spanish Flu pandemic (1918-20)

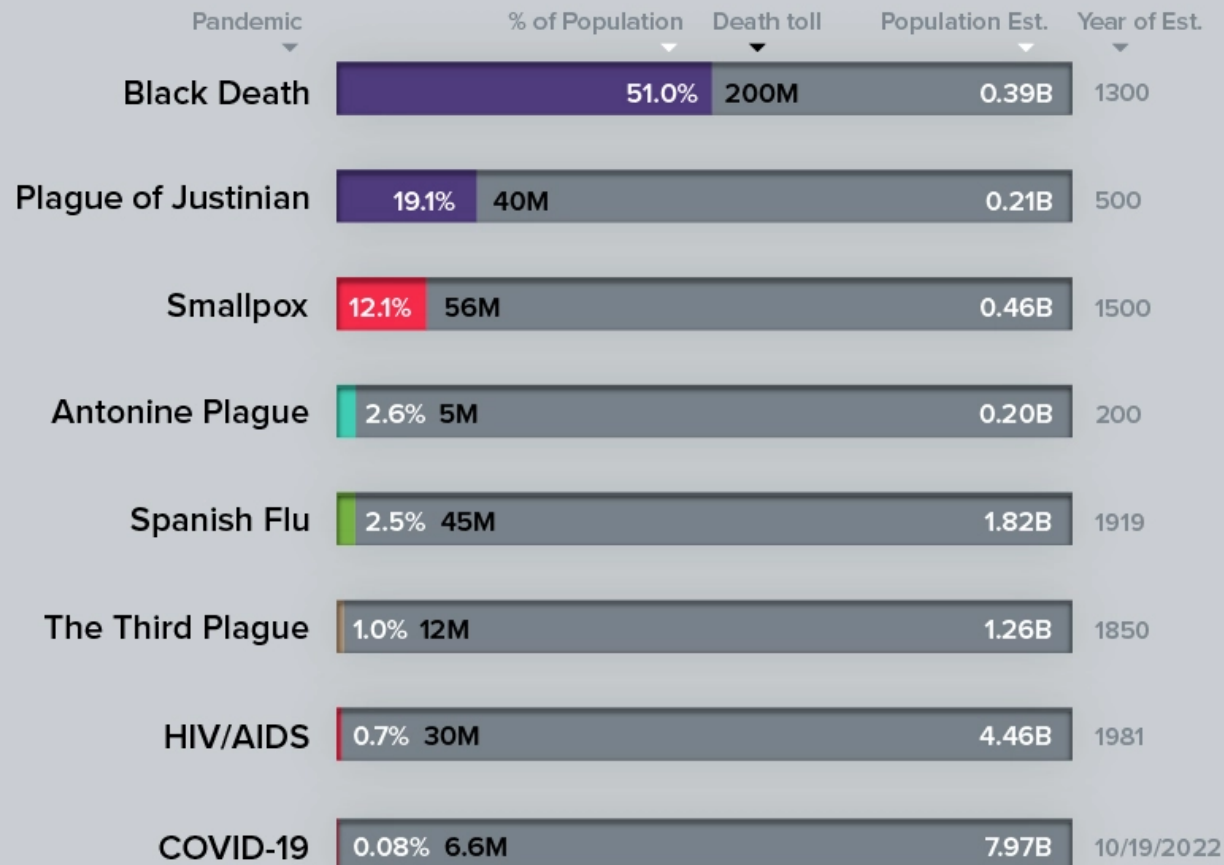
- One of the deadliest pandemics ever

# THE WORLD'S DEADLIEST PANDEMICS

Disease has plagued humanity since the early days of civilization. While outbreaks are a constant issue even in modern times, only a handful of viruses reach full-blown pandemic status.

Here's a look at the deadliest pandemics in history, and their death toll in relation to the global population at the time.

[DEATH TOLL AS A PERCENT OF THE POPULATION]



<https://www.visualcapitalist.com/history-of-pandemics-deadliest/>



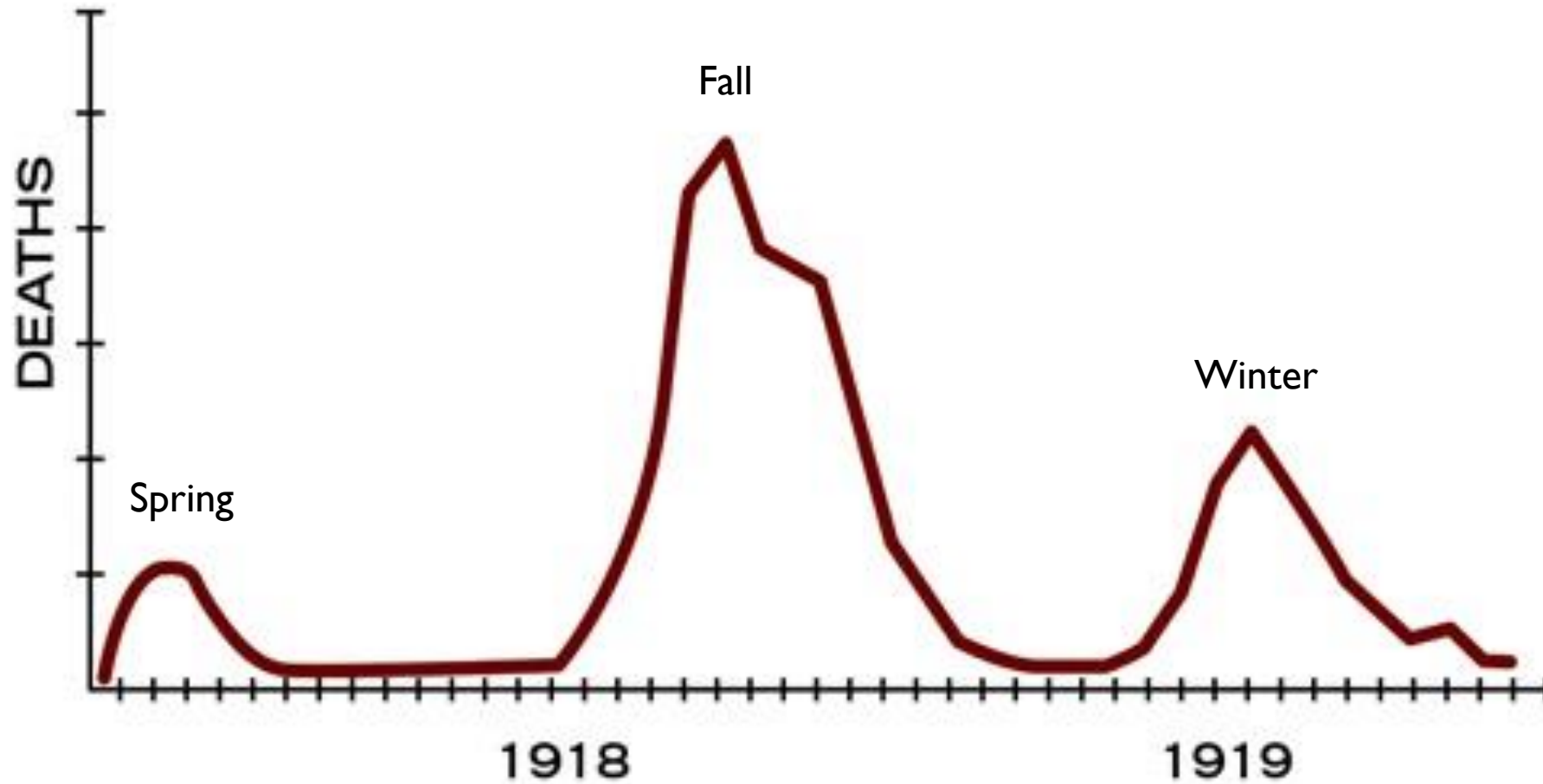
Physics class, University of Montana, Missoula, 1919

<https://www.theatlantic.com/photo/2018/04/photos-the-1918-flu-pandemic/557663/>

# The Spanish Flu pandemic

- One of the deadliest pandemics ever
- Three waves occurred:
  - March 1918: started in the US, and spread to Europe and the rest of the world
  - August 1918: started in France, and spread rapidly to the rest of the world; coincided with the end of WWI
  - Beginning of 1919: some countries were hit by a third wave
- The second wave was the most deadly
- The world death toll ranged between 20 and 45 million people (or more)

# Three waves of the Spanish Flu pandemic in the US



<https://www.cdc.gov/flu/pandemic-resources/1918-commemoration/three-waves.htm>

# The Spanish Flu pandemic

- One of the deadliest pandemics ever
- Three waves occurred:
  - March 1918: started in the US, and spread to Europe and the rest of the world
  - August 1918: started in France, and spread rapidly to the rest of the world; coincided with the end of WWI
  - Beginning of 1919: some countries were hit by a third wave
- The second wave was the most deadly
- The world death toll ranged between 20 and 45 million people (or more)
- Most casualties occurred in India (12-20 million) and China (4-9.5 million)
- Many young adults died (W-pattern of mortality)



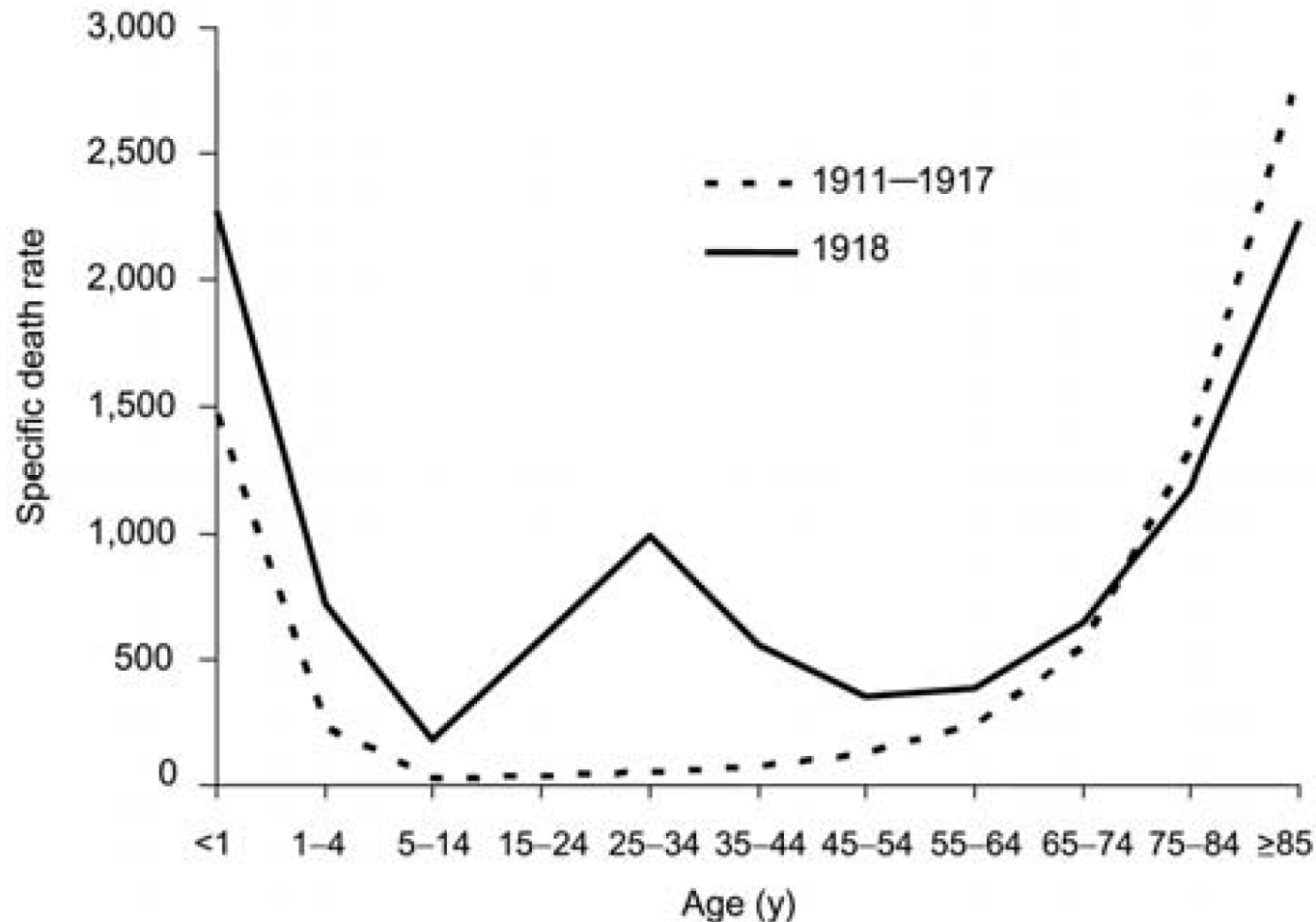


Figure 2. “U-” and “W-” shaped combined influenza and pneumonia mortality, by age at death, per 100,000 persons in each age group, United States, 1911–1918. Influenza- and pneumonia-specific death rates are plotted for the interpandemic years 1911–1917 (dashed line) and for the pandemic year 1918 (solid line) (33,34).

Jeffery K. Taubenberger and David M. Morens (2006), “1918 Influenza: the Mother of All Pandemics”, *Emerging Infectious Diseases*, 12(1): 19.

# Related literature

- Epidemiologists and historians of epidemiology have applied statistical analysis to the Spanish Flu pandemic
- A few examples:
  - Mamelund (2003): mortality among ethnic minorities in Norway
  - Markel et al. (2007): effect of non-pharmaceutical interventions in US cities
  - Mamelund (2011): the influence of geographical isolation
  - Clay, Lewis and Severnini (2019): cross-city variation in 438 US cities
  - Basco, Domènech and Rosés (2021): mortality in Spanish provinces

# PUBLIC HEALTH REPORTS

VOL. 34

AUGUST 8, 1919

No. 32

## INFLUENZA STUDIES.

### I. ON CERTAIN GENERAL STATISTICAL ASPECTS OF THE 1918 EPIDEMIC IN AMERICAN CITIES.<sup>1</sup>

By RAYMOND PEARL, Ph. D., Professor of Biometry and Vital Statistics, School of Hygiene and Public Health, Johns Hopkins University; Consultant in Vital Statistics and Epidemiology, United States Public Health Service.

#### I. Introduction.

The pandemic of influenza which swept over the world in 1918 was the most severe outbreak of this disease which has ever been known, and it takes an unpleasantly high rank in the roster of epidemics generally. It is certainly impossible now, and perhaps always will be, to make any precise statement of the number of people who lost their lives because of this epidemic. But it is certain that the total is an appalling one. Undoubtedly a great many more people died from this cause than from all causes directly connected with the military operations of the Great War. In the United States alone conservative estimates place the deaths from the influenza epidemic at not less than 550,000, which is approximately five times the number (111,179) of American soldiers officially stated<sup>2</sup> to have lost their lives from all causes in the war. And the end of the epidemic is by no means yet reached. In England and Wales the curve of mortality from influenza was even in 1907, seventeen years after the epidemic of 1890, higher than it was in any of the 40 years preceding 1890. The decline in the mortality rate after the 1848 epidemic in Great Britain was similarly slow.<sup>3</sup> There is no evident reason to suppose that conditions following the first explosion of this present epidemic will be essentially different from those which obtained in the earlier cases.

# PUBLIC HEALTH REPORTS

VOL. 36.

FEBRUARY 18, 1921

No. 7

## INFLUENZA STUDIES.

By RAYMOND PEARL, Ph. D., D. Sc., LL. D., Professor of Biometry and Vital Statistics, School of Hygiene and Public Health, Johns Hopkins University; Consultant in Vital Statistics and Epidemiology, United States Public Health Service.

### II. FURTHER DATA ON THE CORRELATION OF EXPLOSIVENESS OF OUTBREAK OF THE 1918 EPIDEMIC.<sup>1</sup>

#### I. Introduction.

In the first of these Studies <sup>2</sup> it was shown that there was a definite and sensible net correlation between explosiveness of outbreak of the epidemic, as measured by an epidemicity index, and the normal death rate from certain organic and chronic diseases. Because of the importance of the subject it has been thought desirable to reexamine critically the data, making use of more refined quantitative measures of the several variables dealt with. It is the object of the present paper to give the results of this re-study of the problem. As before, the basic data are from the large American cities for which weekly data were furnished during the epidemic, by the Bureau of the Census.<sup>3</sup> There is now in progress in this laboratory an extended study of the same problems on the basis of data from the 96 great towns of England and Wales, as well as further studies on the American data.

Before taking up the detailed matters of the present study, I should like to call attention briefly to some methodological considerations which lie at the foundation of this and other papers in this series. It is hoped that in this way the nonmathematical reader may gain a more adequate conception of the real meaning of the results.

TABLE XV.—Data for correlation of demographic characteristics of cities with explosiveness of epidemic influenza mortality.

# Data for Pearl I (1919) – 39 US cities

## Demographics

City.	Response				
	Epi- demicity. Index $I_5$ .	Density of popu- lation (persons per acre).	Geo- graphical position.	Age distribu- tion $\chi^2$ .	Growth in popu- lation.
Albany.....	13.81	8.89	128	4.76	6.5
Atlanta.....	.92	11.42	920	13.06	72.3
Baltimore.....	18.61	30.57	348	6.81	9.7
Birmingham.....	2.41	5.68	1,028	15.80	245.4
Boston.....	9.62	27.36	.....	7.18	19.6
Buffalo.....	10.55	18.97	376	8.86	20.2
Cambridge.....	7.94	28.23	3	6.51	14.1
Chicago.....	6.61	20.28	828	11.45	28.7
Cincinnati.....	2.15	9.10	712	6.73	11.6
Cleveland.....	4.09	20.08	532	11.88	46.9
Columbus.....	2.74	15.18	616	8.35	44.6
Dayton.....	7.20	12.65	684	6.56	36.6
Fall River.....	11.92	5.91	45	10.87	13.8
Grand Rapids.....	1.68	11.85	720	6.17	28.6
Indianapolis.....	2.15	10.96	776	7.23	38.1
Louisville.....	3.07	16.61	796	7.57	9.4
Los Angeles.....	2.00	2.40	2,520	7.67	211.5
Lowell.....	10.58	13.63	23	7.35	11.9
Memphis.....	8.60	12.06	1,104	14.24	28.1
Milwaukee.....	1.53	26.92	832	10.33	31.0
Minneapolis.....	1.12	11.27	1,084	11.46	48.7
Nashville.....	13.83	10.11	924	9.19	36.5
Newark.....	2.81	27.52	192	10.19	41.2
New Haven.....	3.16	13.06	100	6.81	23.7
New Orleans.....	14.60	2.96	1,332	9.25	18.1
New York.....	5.67	29.54	164	11.79	38.7
Oakland.....	3.35	6.41	2,604	6.51	124.3
Omaha.....	2.91	8.34	1,248	10.83	21.0
Philadelphia.....	20.51	21.02	260	7.19	19.7
Pittsburg.....	7.82	22.81	456	11.53	18.2
Providence.....	5.60	22.35	40	6.88	27.8
Richmond.....	13.91	10.76	460	10.55	50.1
Rochester.....	2.62	18.62	328	6.97	34.2
St. Louis.....	2.11	19.36	1,004	9.51	19.4
St. Paul.....	1.43	7.40	1,072	12.70	31.7
San Francisco.....	4.49	17.55	2,624	12.65	21.6
Syracuse.....	8.97	13.34	248	6.21	26.6
Toledo.....	5.95	10.91	620	7.26	27.8
Washington.....	15.34	9.55	376	6.58	18.8

TABLE XVII.—Data for correlation of explosiveness of influenza epidemic mortality, with death rates from various causes for 1916.

City.	Epi- demic- ity index <i>I</i> <sub>5</sub> .	Death rate from all causes per 1,000.	Death rates per 100,000 living, from—							
			Pulmo- nary tuber- culosis.	Organic heart dis- eases.	Acute nephri- tis and Bright's disease.	Influ- enza.	Pneu- monia (all forms).	Ty- phoid fever.	Cancer.	Meas- les.
Albany.....	13.81	19.3	208.5	235.8	197.2	35.8	161.3	7.5	120.8	24.5
Atlanta.....	.92	15.3	117.0	110.2	158.5	14.7	141.2	22.0	63.5	1.6
Baltimore.....	18.61	18.1	200.5	193.2	174.3	21.5	235.7	18.1	106.7	5.4
Birmingham.....	2.41	14.1	173.9	84.7	85.8	13.2	137.5	43.5	56.1	.....
Boston.....	9.62	16.9	145.0	220.4	102.6	11.2	210.8	3.4	115.8	14.5
Buffalo.....	10.55	16.1	142.8	170.1	127.0	10.2	166.3	10.9	100.7	15.8
Cambridge.....	7.94	13.5	172.6	191.2	70.8	9.7	159.3	1.8	112.4	7.1
Chicago.....	6.61	14.5	132.8	159.9	107.2	11.7	158.1	5.2	91.3	5.4
Cincinnati.....	2.15	16.4	208.3	202.7	158.8	26.8	145.4	3.2	116.2	15.3
Cleveland.....	4.09	14.8	132.2	119.6	90.9	16.3	182.2	5.3	86.8	8.9
Columbus.....	2.74	15.5	125.2	156.4	90.3	33.5	155.9	13.0	100.5	15.8
Dayton.....	7.20	15.2	121.8	180.8	119.5	18.9	146.2	19.7	114.8	1.6
Fall River.....	11.92	17.0	161.3	158.9	105.9	24.1	243.8	10.9	91.9	30.4
Grand Rapids.....	1.68	12.2	64.7	134.8	88.9	9.4	70.2	16.4	88.1	2.3
Indianapolis.....	2.15	15.6	159.6	175.6	115.0	17.4	141.8	26.1	99.4	9.8
Louisville.....	3.07	15.0	159.9	145.7	154.0	33.1	146.9	13.4	83.7	2.1
Los Angeles.....	2.00	12.3	176.7	161.0	111.3	9.3	78.0	2.6	105.6	2.0
Lowell.....	10.58	17.3	103.3	161.6	89.2	14.1	178.4	11.5	85.7	25.6
Memphis.....	8.60	19.8	262.1	145.1	171.1	37.0	136.9	26.7	83.2	2.7
Milwaukee.....	1.53	12.7	78.8	102.9	79.9	15.8	154.2	15.3	92.8	27.7
Minneapolis.....	1.12	12.4	117.8	120.0	101.8	8.8	111.4	5.5	96.0	20.4
Nashville.....	13.83	17.2	201.8	211.2	132.8	25.0	152.6	37.1	77.6	.9
Newark.....	2.81	15.0	145.5	153.6	140.9	17.4	161.2	6.1	85.6	25.7
New Haven.....	3.16	17.0	95.5	175.0	122.3	37.4	225.1	8.7	116.2	5.3
New Orleans.....	14.60	18.4	259.0	207.4	231.1	26.9	117.3	23.1	93.1	3.5
New York.....	5.67	13.9	154.9	168.7	131.4	9.8	179.9	3.9	84.5	9.9
Oakland.....	3.35	10.5	94.2	189.3	89.1	8.6	75.5	4.0	89.6	.....
Omaha.....	2.91	14.4	101.5	93.7	91.3	18.7	173.4	3.0	90.0	1.8
Philadelphia.....	20.51	16.2	170.6	197.4	177.7	24.0	172.2	7.6	101.1	6.6
Pittsburgh.....	7.82	17.4	110.7	144.7	92.0	26.6	331.0	9.0	89.8	23.7
Providence.....	5.60	15.8	134.1	167.5	142.4	25.9	174.1	5.1	100.0	25.1
Richmond.....	13.91	19.7	187.0	189.5	204.9	20.4	194.0	23.6	97.0	26.2
Rochester.....	2.62	14.4	91.9	192.3	136.7	8.9	121.6	5.0	114.7	8.1
St. Louis.....	2.11	14.9	129.0	144.6	176.8	22.8	173.5	9.4	95.3	8.8
St. Paul.....	1.43	11.3	99.1	122.6	92.6	9.3	80.5	5.7	87.0	7.3
San Francisco.....	4.49	15.4	169.4	250.7	135.3	4.1	129.0	3.5	133.1	1.3
Syracuse.....	8.97	15.2	88.0	201.1	112.5	10.9	134.3	12.2	110.5	.....
Toledo.....	5.95	18.1	168.1	192.8	89.3	19.7	156.5	22.2	97.9	33.8
Washington.....	15.34	17.8	187.4	230.5	168.1	24.2	164.3	12.9	107.7	2.2

# Data for Pearl I (1919) – 39 US cities

## Death rates for 1916

#### IV. Epidemicity Indices.

With the variation data in hand one further step is necessary before the analysis by multiple correlation can be completed. We must have a single numerical measure or index of the force of the epidemic explosion in any particular place. In the earlier sections we have seen that the mortality curves in some cities have a single very sharp peak, while in other cases the curve of epidemic mortality is a long, low, flat curve. To deal practically with such differences, it is essential to have some single numerical index which will be sensitive to changes of any order in the curve, and at the same time will measure the essential characteristic which we want to measure in an epidemic curve.

TABLE XIII.—Showing values of different epidemicity indices of mortality in American cities during influenza epidemic of 1918.

Cities.	$I_1$ (weeks).	$I_2$ (per cent).	$I_3$ .	$I_4$ .	$I_5$ .
Albany.....	1.61	85.9	40.13	4.7	13.81
Atlanta.....	6.68	58.5	9.31	2.7	.92
Baltimore.....	1.54	94.5	48.61	6.1	18.61
Birmingham.....	4.06	60.7	17.04	.....	2.41
Boston.....	1.98	88.5	33.47	6.5	9.62
Buffalo.....	1.85	92.0	31.19	5.8	10.55
Cambridge.....	2.00	88.9	27.68	5.9	7.94
Chicago.....	1.98	72.4	24.04	3.8	6.61
Cincinnati.....	4.55	69.8	15.41	4.0	2.15
Cleveland.....	3.63	74.2	18.30	4.0	4.09
Columbus.....	3.55	56.4	14.94	3.2	2.74
Dayton.....	6.24	91.4	24.67	3.5	7.20
Fall River.....	1.66	80.9	38.70	5.8	11.92
Grand Rapids.....	3.41	65.7	8.10	1.5	1.68
Indianapolis.....	3.42	55.9	12.51	2.5	2.15
Louisville.....	4.11	78.4	15.45	3.6	3.07
Los Angeles.....	5.50	62.7	15.78	5.2	2.00
Lowell.....	1.70	71.5	34.60	5.1	10.58
Memphis.....	1.76	94.7	24.15	.....	8.60
Milwaukee.....	4.48	57.4	11.57	2.9	1.53
Minneapolis.....	5.98	55.1	9.80	2.7	1.12
Nashville.....	1.58	72.6	39.39	7.8	13.83
Newark.....	5.70	99.0	15.34	5.1	2.81
New Haven.....	5.43	100.6	18.89	5.6	3.16
New Orleans.....	1.69	90.2	40.95	7.2	14.60
New York.....	2.19	71.2	23.29	4.7	5.67
Oakland.....	5.25	77.0	18.74	5.9	3.35
Omaha.....	4.17	69.6	18.47	.....	2.91
Philadelphia.....	1.52	86.2	56.08	7.3	20.51
Pittsburgh.....	2.79	67.0	37.62	8.0	7.82
Providence.....	2.46	86.4	21.79	5.3	5.60
Richmond.....	1.33	66.1	35.12	.....	13.91
Rochester.....	4.48	79.2	13.94	2.7	2.62
St. Louis.....	4.06	59.1	13.47	3.0	2.11
St. Paul.....	5.12	57.8	11.31	3.3	1.43
San Francisco.....	5.06	78.4	26.50	7.5	4.49
Syracuse.....	2.09	94.2	30.77	.....	8.97
Toledo.....	1.67	69.8	17.19	2.1	5.95
Washington.....	1.49	66.3	45.08	6.6	15.34

# Updated variables in Pearl II (1921) – 34 US cities

Subscript No.

Variable.

2 responses

1. Explosiveness of outbreak of epidemic mortality as measured by an epidemicity index  $I_6$ . + Excess mortality rate as measured by the “destructiveness” variable
- 3a. Normal death rate from pulmonary tuberculosis.
- 3b. Normal death rate from organic diseases of the heart.
- 3c. Normal death rate from acute nephritis and Bright’s disease.
- 3d. Normal death rate from typhoid fever.
- 3e. Normal death rate from cancer and other malignant tumors.
- 3f. Normal death rate from all causes.
4. Age distribution of population.
5. Sex ratio of population.
6. Density of population.
7. Latitude.
8. Longitude.
9. Rate of growth of population, 1900-1910.

*Death rates averaged over 1915-1917*

*Demographics*



# Additional data

- ▲ Persons to a dwelling
- ▲ Percentage of homes owned
- ▲ School attendance of population 6 to 20 years of age
- ▲ Illiteracy in the population 10 years of age and over
- ▲ Share ages 0-4
- ▲ Share ages 5-14
- ▲ Share ages 15-24
- ▲ Share ages 25-44
- ▲ Share ages 45-64
- ▲ Share ages 65 -

## 1910

### Information about the 1910 Census

Under the provisions of the census act of July 2, 1909, the thirteenth census was administered. In accordance with the provisions of the act, general population and Indian population schedules were prepared. The schedules used for Hawaii and Puerto Rico, although similar to the general population schedule, differed slightly from those used within the United States.

Census enumerators began canvassing the Nation on April 15, 1910.<sup>1</sup> The law gave census takers 2 weeks to complete their work in cities of 5,000 inhabitants or more, while enumerators in smaller and rural areas were allotted 30 days to complete their task.

<sup>1</sup> The change of “census day” from June 1 to April 15 was made up on the suggestion of the Census Bureau. It was believed that the April 15 date would be more desirable, since a large number of people are away from their homes in June.



Partners Researchers Educators Survey Respondents

News NAICS Codes Jobs About Us Contact Us Help

Topics Data & Maps **Surveys & Programs** Resource Library

Search data, events, resources, and more 🔍

// [Census.gov](#) / [Our Surveys & Programs](#) / [Decennial Census of Population and Housing](#) / [By Decade](#) / [Decennial Census Official Publications](#)

Within Decennial Census of  
Population and Housing

[About](#)

[By Decade](#)

## Decennial Census Official Publications

<https://www.census.gov/programs-surveys/decennial-census/decade/decennial-publications.1910.html>

# Pearl's analysis of (partial) correlations

## VI. The Correlation of the Explosiveness of the Outbreak of Mortality in the Influenza Epidemic with Various Other Factors.

We come now to the most essential part of the study, namely, the attempt to find factors directly related to or concerned in the production of the extraordinary differences between different cities in respect of the relative explosiveness of the outbreak of epidemic mortality. The method of analysis which will be followed is that of multiple correlation.<sup>1</sup> The general principle of the correlation method is simple. If in the present case, for example, we should find that, in general, when a city had a high influenza epidemicity index it also had a high density of population, and conversely, that cities having low epidemicity indices had low density of population, it would be said that there was a positive correlation in variation between explosiveness of epidemic and density of population.

In a system of  $n$  variables correlation between any two, with the others remaining constant, is measured by the coefficient.

$$r_{12.34 \dots n} = \frac{r_{12.34 \dots (n-1)} - r_{1n.34 \dots (n-1)} \cdot r_{2n.34 \dots (n-1)}}{(1 - r_{1n.34 \dots (n-1)}^2)^{\frac{1}{2}} (1 - r_{2n.34 \dots (n-1)}^2)^{\frac{1}{2}}}$$

and a coefficient of zero order is found from the observations by the following well-known expression:

$$r_{12} = \frac{S(xy)}{N\sigma_1\sigma_2}$$

# Variable selection with null factor and bootstrap simulation

Controlling the type-I error rate using Wu, Boos and Stefanski (JASA, 2007):

## Controlling Variable Selection by the Addition of Pseudovariabes

Yujun WU, Dennis D. BOOS, and Leonard A. STEFANSKI

---

We propose a new approach to variable selection designed to control the false selection rate (FSR), defined as the proportion of uninformative variables included in selected models. The method works by adding a known number of pseudovariabes to the real dataset, running a variable selection procedure, and monitoring the proportion of pseudovariabes falsely selected. Information obtained from bootstrap-like replications of this process is used to estimate the proportion of falsely selected real variables and to tune the selection procedure to control the FSR.

KEY WORDS: False selection rate; Forward selection; Model error; Model selection; Subset selection.

---

# Null factor and bootstrap simulation

- We included **a single null factor** in the model and performed 2500 bootstrap replicates for variable selection using JMP
- We calculated the proportion of times each variable enters the model
- Variables that enter as often or less than the null factor are ignorable

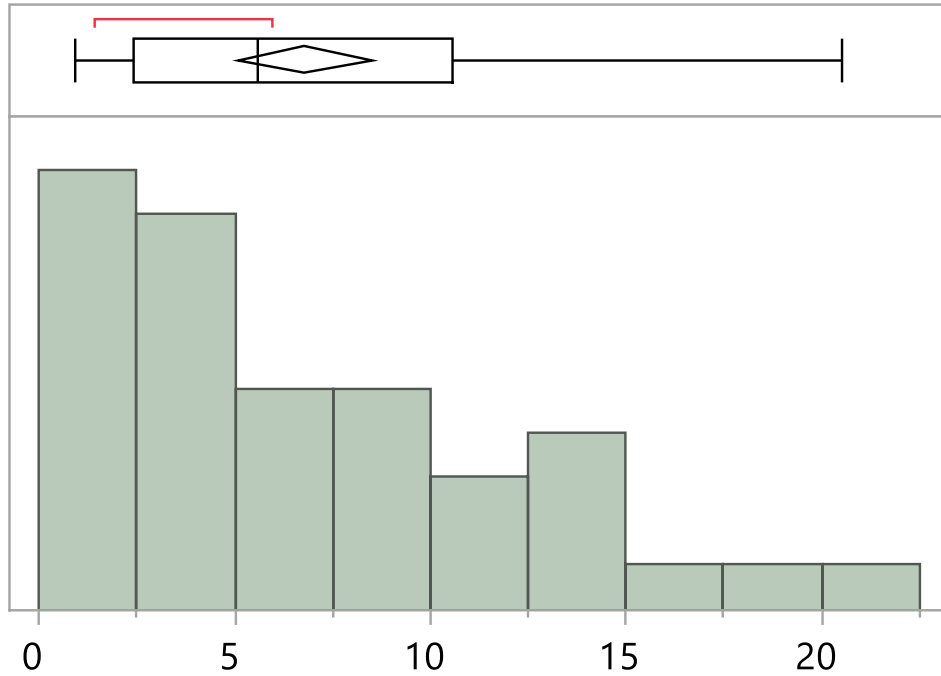
# Null factor and bootstrap simulation

Bootstrap frequency column

```
Null Factor Row ( ) = Random Normal ( <mu=0> );  
Resample Freq ( <rate> , <column> );
```

Frq	Null Factor
0	-1.937037...
0	0.492011...
0	-0.949424...
1	-0.067927...
0	-0.130842...
2	-0.906760...
0	0.068270...
2	-0.433499...
1	1.984110...
2	-0.825345...

## Epidemicity Index I5



### Summary Statistics

Mean	6.7789744
Std Dev	5.2833337
Std Err Mean	0.8460105
Upper 95% Mean	8.491633
Lower 95% Mean	5.0663157
N	39
N Missing	0

## Poisson Goodness of Fit test in the Generalized Linear Model platform

### Generalized Linear Model Fit

Freq: Frq

Response: Epidemicity Index I5

Distribution: Poisson

Link: Log

Estimation Method: Maximum Likelihood

Observations (or Sum Wgts) = 39

### Whole Model Test

Model	-LogLikelihood	ChiSquare	DF	Prob>ChiSq
Difference	53.6332308	107.2665	14	<.0001 *
Full	74.316988			
Reduced	127.950219			

### Goodness Of

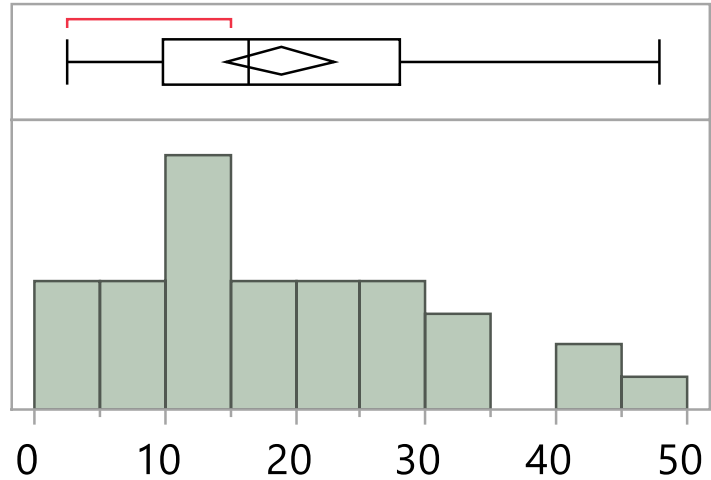
Fit Statistic	ChiSquare	DF	Prob>ChiSq
Pearson	18.2971	24	0.7884
Deviance	18.7681	24	0.7642

### AICc

199.5035

## Poisson regression

### Epidemicity Index I6

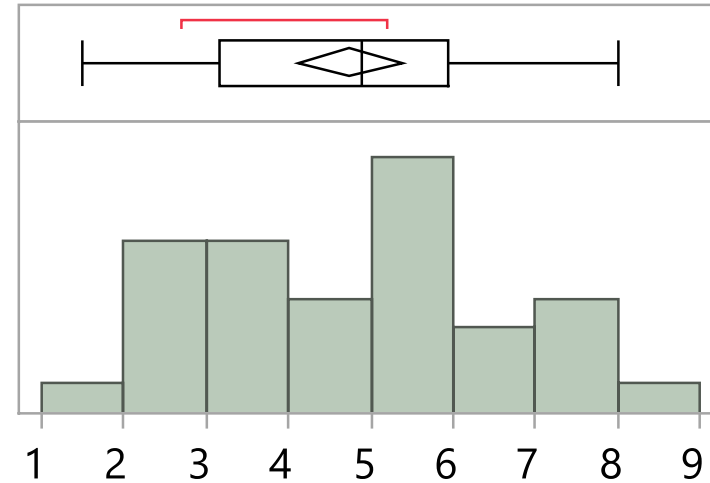


### Summary Statistics

Mean	18.894118
Std Dev	11.919552
Std Err Mean	2.0441863
Upper 95% Mean	23.053046
Lower 95% Mean	14.735189
N	34
N Missing	0

## Normal regression

### Destructiveness



### Summary Statistics

Mean	4.7529412
Std Dev	1.7899103
Std Err Mean	0.3069671
Upper 95% Mean	5.3774704
Lower 95% Mean	4.1284119
N	34
N Missing	0

### Model Specification

#### Select Columns

30 Columns

Enter column name

- US City
- Epidemicity Index I5
- Population Density
- Geographical Position
- Age Distribution
- Perc Pop growth
- DR All Causes
- DR Pulmonary Tuberculosis
- DR Organic Heart Disease
- DR Acute Nephritis N Bright's Disease
- DR Influenza
- DR Pneumonia
- DR Typhoid Fever
- DR Cancer
- DR Measles

#### Pick Role Variables

Y	Epidemicity Index I5 <i>optional</i>
Weight	<i>optional numeric</i>
Freq	Frq
Validation	<i>optional numeric</i>
By	<i>optional</i>

Personality: Generalized Regression

Distribution: Poisson

Help Run

Recall  Keep dialog open

Remove

#### Construct Model Effects

Add	Population Density
Cross	Geographical Position
Nest	Age Distribution
Macros ▾	Perc Pop growth
Degree <input type="text" value="2"/>	DR All Causes
Attributes ▾	DR Pulmonary Tuberculosis
Transform ▾	DR Organic Heart Disease
<input type="checkbox"/> No Intercept	DR Acute Nephritis N Bright's Disease
	DR Influenza
	DR Pneumonia



# Forward variable selection

## Generalized Regression for Epidemicity Index I5

### Model Launch

Response Distribution

Poisson

Estimation Method

Forward Selection

Validation Method

AICc

Early Stopping

### Poisson Forward Selection with AICc Validation

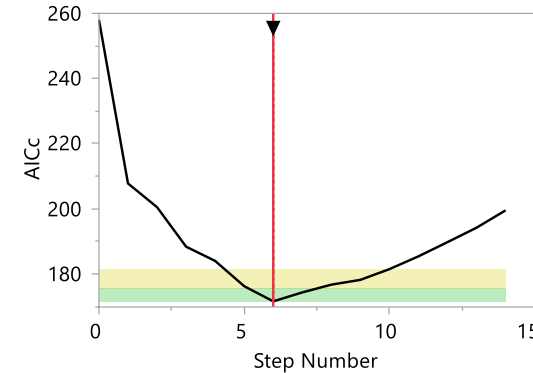
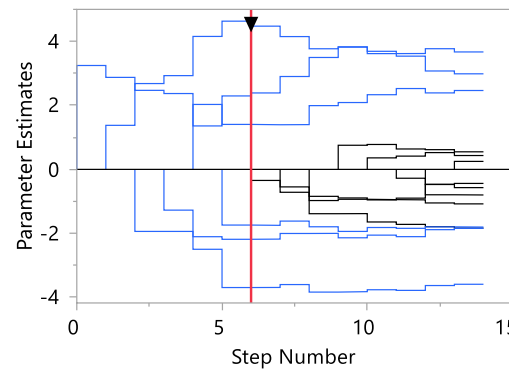
#### Model Summary

Response Epidemicity Index I5  
 Distribution Poisson  
 Estimation Method Forward Selection  
 Validation Method AICc  
 Mean Model Link Log

#### Measure

Number of rows 39  
 Sum of Frequencies 39  
 -LogLikelihood 76.968437  
 Number of Parameters 7  
 BIC 179.58181  
 AICc 171.54978  
 Generalized RSquare 0.926792

#### Solution Path



#### Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald		Prob >	
			ChiSquare	ChiSquare	Lower 95%	Upper 95%
Intercept	-0.636276	0.8018115	0.6297195	0.4275	-2.207798	0.9352453
Population Density	0	0	0	1.0000	0	0
Geographical Position	0	0	0	1.0000	0	0
Age Distribution	0	0	0	1.0000	0	0
Perc Pop growth	0	0	0	1.0000	0	0
DR All Causes	0.1755972	0.0538666	10.626645	0.0011 *	0.0700206	0.2811739
DR Pulmonary Tuberculosis	0	0	0	1.0000	0	0
DR Organic Heart Disease	0.0192042	0.0030112	40.673334	<.0001 *	0.0133023	0.025106
DR Acute Nephritis N Bright's Disease	0	0	0	1.0000	0	0
DR Influenza	0	0	0	1.0000	0	0
DR Pneumonia	0.0038551	0.0018386	4.3965342	0.0360 *	0.0002516	0.0074586
DR Typhoid Fever	-0.029401	0.0110204	7.1173173	0.0076 *	-0.051	-0.007801
DR Cancer	-0.040974	0.0081518	25.264359	<.0001 *	-0.056951	-0.024997
DR Measles	0	0	0	1.0000	0	0
Null Factor	-0.391098	0.1032916	14.336409	0.0002 *	-0.593545	-0.18865

# Bootstrap simulation

## Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	276	0.8018115	0.6297195	0.4275	-2.207798	0.9352453
Population Density	0	0	0	1.0000	0	0
Geographical Position	0	0	0	1.0000	0	0
Age Distribution	0	0	0	1.0000	0	0
Perc Pop growth	0	0	0	1.0000	0	0
DR All Causes	972	0.0538666	10.626645	0.0011*	0.0700206	0.2811739
DR Pulmonary	0	0	0	1.0000	0	0
DR Organic Heart Disease	042	0.0030112	40.673334	<.0001*	0.0133023	0.025106
DR Acute Nephritis	0	0	0	1.0000	0	0
DR Influenza	0	0	0	1.0000	0	0
DR Pneumonia	551	0.0018386	4.3965342	0.0360*	0.0002516	0.0074586
DR Typhoid Fever	401	0.0110204	7.1173173	0.0076*	-0.051	-0.007801
	974	0.0081518	25.264359	<.0001*	-0.056951	-0.024997
Null Factor	-0.391098	0.1032916	14.336409	0.0002*	-0.593545	-0.18865

- Table Style
- Columns
- Sort by Column...
- Make into Data Table
- Make Combined Data Table
- Make Into Matrix
- Select Where...
- Filter Where...
- Format Column...
- Align Decimal Separator
- Show Properties
- Copy Column
- Copy Table
- Simulate
- Bootstrap

Switch columns to perform a simulation.

Simulation

Column to Switch Out

- Frq
- Epidemicity Index I5
- Population Density
- Geograp...l Position
- Age Distribution
- Perc Pop growth
- DR All Causes
- DR Pulm...berculosis
- DR Orga...rt Disease
- DR Acute...s Disease
- DR Influenza
- DR Pneumonia
- DR Typhoid Fever
- DR Cancer
- DR Measles

Column to Switch In

- Frq

Number of Samples

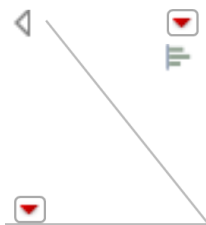
Random Seed

# Selected variables

	Factor	Proportion Nonzero In Bootstrap Simulation (FS Poisson AICc)	NF 99.9% (Sim) Upper CL
1	DR All Causes	0.828	0.439
2	DR Organic Heart Disease	0.804	0.439
3	DR Pneumonia	0.524	0.439
4	DR Cancer	0.512	0.439
5	DR Measles	0.501	0.439
6	Geographical Position	0.492	0.439
7	Null Factor	0.410	0.439
8	DR Influenza	0.399	0.439
9	Population Density	0.370	0.439
10	DR Pulmonary Tuberculosis	0.272	0.439
11	Age Distribution	0.245	0.439
12	DR Acute Nephritis N Bright's Disease	0.217	0.439
13	DR Typhoid Fever	0.194	0.439
14	Perc Pop growth	0.110	0.439

# Selected variables (without DR All Causes)

Highly correlated with DR All Causes



	Factor	Proportion Nonzero In Bootstrap Simulation (FS Poisson AICc)	NF 99.9% (Sim) Upper CL
1	DR Organic Heart Disease	0.948	0.427
2	DR Pneumonia	0.916	0.427
3	Geographical Position	0.562	0.427
4	DR Pulmonary Tuberculosis	0.527	0.427
5	DR Typhoid Fever	0.444	0.427
6	DR Cancer	0.430	0.427
7	DR Acute Nephritis N Bright's Disease	0.423	0.427
8	Null Factor	0.398	0.427
9	DR Influenza	0.336	0.427
10	DR Measles	0.268	0.427
11	Age Distribution	0.197	0.427
12	Population Density	0.192	0.427
13	Perc Pop growth	0.156	0.427

- Poisson Forward Selection model for  $I_5$

Parameter Estimates for Original Predictors		$e^{0.01486} = 1.0149$				
Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
DR Organic Heart Disease	0.0148618	0.0026863	30.606827	<.0001 *	0.0095966	0.0201269
Share ages 0-4	0.2629124	0.0785037	11.216092	0.0008 *	0.1090479	0.4167768
School attendance of population 6 to 20 years of age	-0.049186	0.0181879	7.3133069	0.0068 *	-0.084834	-0.013538
DR All Causes	0.0932252	0.0486634	3.6699573	0.0554	-0.002153	0.1886037
Geographical Position	-0.000262	0.0001526	2.9556235	0.0856	-0.000561	3.6743e-5
Intercept	-1.567006	1.9082925	0.6742972	0.4116	-5.30719	2.1731787

- Pearl I

TABLE XVIII.—Mean and standard deviation for death rates from various causes.

Cause of death.	Mean death rate.	Standard deviation in death rate.	Coefficient of correlation between epidemicity index $I_5$ and the death rate from the specified cause.
All causes <sup>1</sup> .....	15.55±0.24	2.21±0.17	+0.661±0.061
Pulmonary tuberculosis.....	147.50±4.94	45.73±3.49	+ .525± .078
Organic heart disease.....	168.29±4.19	38.82±2.96	+ .567± .073
Acute nephritis and Bright's disease.....	127.39±4.17	38.57±2.95	+ .507± .080
Influenza.....	18.80± .96	8.86± .68	+ .287± .099
Pneumonia (all forms).....	158.40±5.18	47.99±3.66	+ .388± .092
Typhoid fever.....	12.41±1.04	9.64± .74	+ .176± .105
Cancer.....	97.07±1.62	14.99±1.14	+ .198± .104
Measles.....	11.00±1.09	10.08± .77	+ .069± .107

<sup>1</sup> Death rate per 1,000; in all other cases in the table the death rate is per 100,000.

- Poisson Forward Selection model for  $I_6$

Parameter Estimates for Original Predictors						
Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
DR Organic Heart Disease	0.0107234	0.0018892	32.219029	<.0001 *	0.0070207	0.0144262
Intercept	-2.485853	0.5943956	17.490393	<.0001 *	-3.650847	-1.320859
DR All Causes	0.0011181	0.0002758	16.433068	<.0001 *	0.0005775	0.0016587
Share ages 0-4	0.1651742	0.0464375	12.65162	0.0004 *	0.0741583	0.2561901
Population Density	0.0124739	0.0053104	5.5176314	0.0188 *	0.0020657	0.0228821

- Pearl II

TABLE III.—*Net correlation of explosiveness of outbreak ( $I_6$ ) with the normal death rates from certain specified causes.*

Variable correlated with explosiveness ( $I_6$ ): Death rate from—	$r$ subscripts.	Coefficient.
All causes.....	13f. 456789	+0.572 ±0.078
Pneumonia.....	13a. 456789	+0.389 ±0.098
Organic diseases of the heart.....	13b. 456789	+0.562 ±0.079
Acute nephritis and Bright's disease.....	13c. 456789	+0.307 ±0.105
Typhoid fever.....	13d. 456789	+0.105 ±0.114
Cancer and other malignant tumors.....	13e. 456789	+0.141 ±0.113

- Normal Forward Selection model for *Destructiveness*

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-20.86081	9.31124	-2.24	0.0326 *
DR Organic Heart Disease	0.0324243	0.007641	4.24	0.0002 *
Share ages 0-4	0.9657136	0.339097	2.85	0.0079 *
Share ages 25-44	0.3227745	0.18789	1.72	0.0961

- Pearl II

TABLE II.—*Net correlation of destructiveness (25-week excess mortality) with the normal death rate from certain specified causes.*

Variable correlated with destructiveness (25-week excess mortality). Death rate from—	r subscripts.	Coefficient.
All causes.....	23f. 456789	+0.405±0.097
Pulmonary tuberculosis.....	23a. 456789	+0.279±0.107
Organic diseases of the heart.....	23b. 456789	+0.537±0.082
Nephritis and Bright's disease.....	23c. 456789	-0.098±0.116
Typhoid fever.....	23d. 456789	-0.133±0.113
Cancer and other malignant tumors.....	23e. 456789	+0.268±0.107

# Conclusion

- Pearl's data sets are very tiny and observational with multicollinearity
- George Box: "All models are wrong, but some are useful"
  - Pearl's correlation analysis and our null factor analysis are useful, but not magical
  - Pearl's first analysis is not fully supported (and he knew it !)
  - We selected *satisfactory models* in a sequential manner:
    1. We included Pearl's variables
    2. We retained the selected variables
    3. We included Pearl's selected variables from 2. together with new 1910 Census variables
    4. We retained all selected variables