社会统计学及SPSS软件应用
STATISTICS WITH SPSS

Instructor:王荣欣

Email: rxwang@qq.com

周二3-4节、单周周四3-4节，3A106-2

2020年9月21日

CONTENTS

CONTENTS

概率论基本知识

1 概率
2 随机事件
3 随机变量、随机变量的期望

概率论基本知识

1. probability function p(x)
2. the expected value of X corresponding to some probability function p(x) is

$$E(X) = \sum xp(x)$$

$$\mu = E(X)$$

概率论基本知识

1. population mean μ
2. population variance

$$\sigma^2 = E[(X-\mu)^2] = \sum (X-\mu)^2 f(x)$$

3. population standard deviation σ

概率论基本知识

- discrete distribution:
  - binomial probability function
- probabilities associated with continuous variables are given by the area under a curve. The equation for this curve is called probability density function.

连续变量的常用分布

1. normal distribution
2. chi-square distribution
3. F distribution
4. t distribution

DESCRIPTIVE STATISTICS

- 单个变量
  1. 集中趋势
  2. 离散趋势
- 两个变量 （cross-tabulation）
  1. 两个定类变量
  2. 两个定序变量
  3. 两个定距变量
  4. 一个定类、一个定序
  5. 一个定类、一个定距
  6. 一个定序、一个定距

单个变量

- statistical measures of central tendency
  1. mean
  2. median
  3. mode
- statistical measures of variation
  1. quartile and percentile
  2. range
  3. standard deviation

MEASURES OF CENTRAL TENDENCY

1. The mean is what most people are accustomed to calling an average.
2. It involves adding up all of the values in a set of data (X) and then dividing by the total number (N) of cases.

MEASURES OF CENTRAL TENDENCY

- The median is the point that divides the distribution into a lower half and an upper half so that 50% of the values are in one half and 50% are in the other.
- Francis Galton wanted to find a faster way to establish an average, rather than going through the trouble of calculating the mean value. He introduced the word percentile, which is the point that divides a distribution into a lower percentage of cases and an upper percentage.

MEASURES OF CENTRAL TENDENCY

Through Gauss first used the median in 1816, Galton introduced it into statistics. In 1874 he devised a statistical scale to find the median when he introduced the 50th percentile as the middle point in a set of data, where a set of data is divided exactly into half.

MEASURES OF CENTRAL TENDENCY

- The mode, quoted by Karl Pearson in 1894, is the value that occurs more frequently than any other. This finds its greatest use in advertising, which deals in concepts like the ”modal family”.
- The mode is a point of maximum frequency; it is used most often to look for typical cases. The mode may or may not compare to an actual value.

MEASURES OF CENTRAL TENDENCY

- The mode, quoted by Karl Pearson in 1894, is the value that occurs more frequently than any other. This finds its greatest use in advertising, which deals in concepts like the ”modal family”.
- The mode is a point of maximum frequency; it is used most often to look for typical cases. The mode may or may not compare to an actual value.

STATISTICAL MEASURES OF VARIATION

- Galton devised the first measure of statistical variation in 1875 when he introduced the semi-interquartile range, which he expressed as:

$$\frac{Q3 - Q1}{2}$$

A quartile is a point on the distribution.

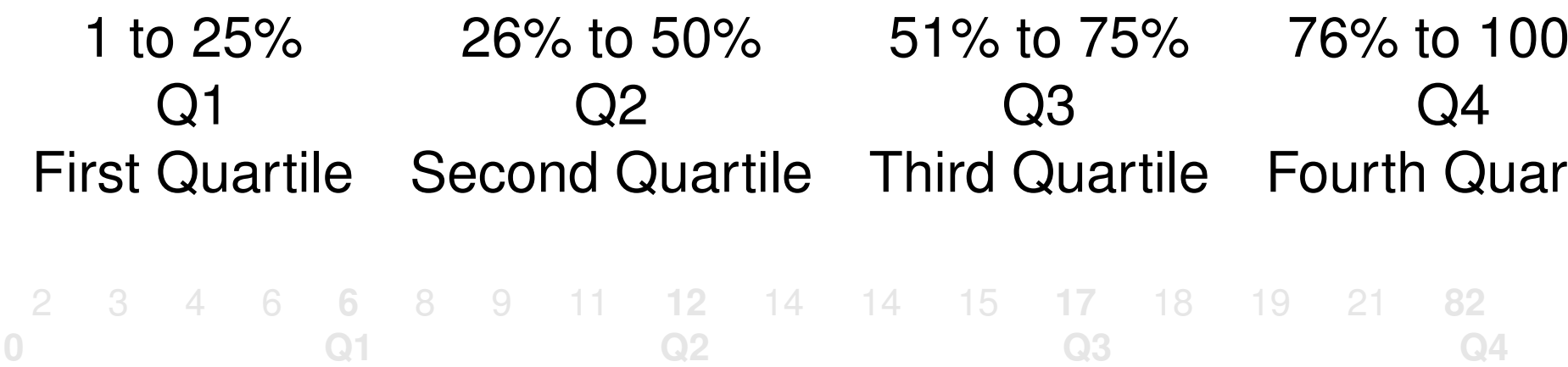
STATISTICAL MEASURES OF VARIATION

- Galton devised the first measure of statistical variation in 1875 when he introduced the semi-interquartile range, which he expressed as:

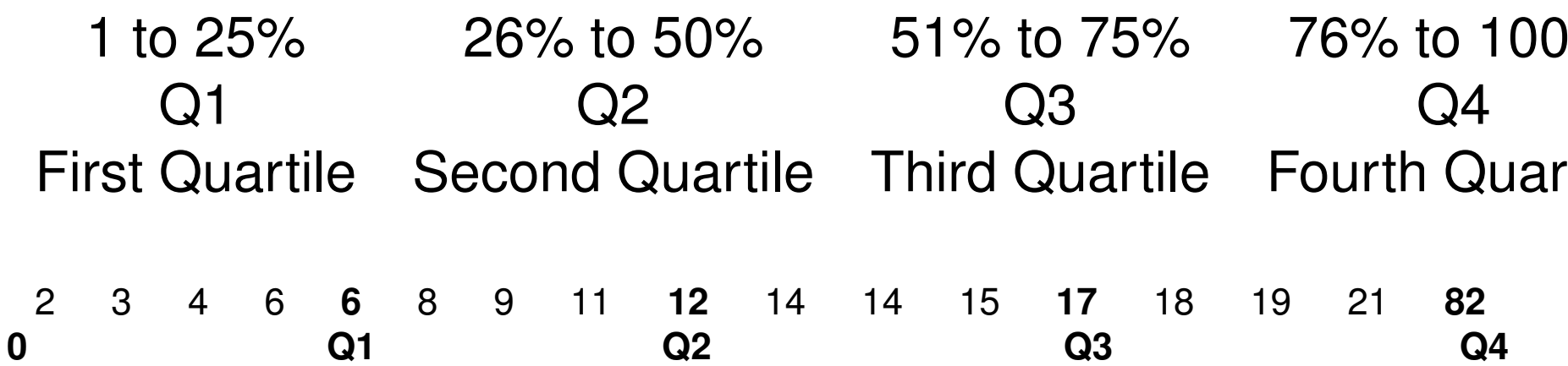
$$\frac{Q3 - Q1}{2}$$

A quartile is a point on the distribution.

STATISTICAL MEASURES OF VARIATION



STATISTICAL MEASURES OF VARIATION



STATISTICAL MEASURES OF VARIATION

- In 1892, Pearson introduced the range, which is the simplest method used to measure variation. The range measures the distance between the largest and smallest values from a particular set of measurements and gives an idea of the spread of the data.
- It is quite often used for summaries of data made available to the general public, such as the range of salaries, ages and temperatures.

STATISTICAL MEASURES OF VARIATION

- In 1892, Pearson introduced the range, which is the simplest method used to measure variation. The range measures the distance between the largest and smallest values from a particular set of measurements and gives an idea of the spread of the data.
- It is quite often used for summaries of data made available to the general public, such as the range of salaries, ages and temperatures.

STATISTICAL MEASURES OF VARIATION

- Pearson introduced the standard deviation in 1893, referring to it initially as the ”standard divergence”. The standard deviation is a measure of variation. It indicates how widely or closely spread the values are in a set of a data, and shows how much each of these individual values deviate from the average (i.e. the mean)
- The standard deviation (σ) corresponds to the ”moment of inertia” and the covariance ($\sum xy$) corresponds to the ”product-moment of dynamics” (the moment of dynamics is concerned with the effect of force on the motion of objects).

STATISTICAL MEASURES OF VARIATION

- Pearson introduced the standard deviation in 1893, referring to it initially as the ”standard divergence”. The standard deviation is a measure of variation. It indicates how widely or closely spread the values are in a set of a data, and shows how much each of these individual values deviate from the average (i.e. the mean)
- The standard deviation (σ) corresponds to the ”moment of inertia” and the covariance ($\sum xy$) corresponds to the ”product-moment of dynamics” (the moment of dynamics is concerned with the effect of force on the motion of objects).

STATISTICAL MEASURES OF VARIATION

The covariance is the measure of how much two random variables move together.

1. If two variables tend to move together in the same direction, then the covariance between the two variables will be positive.
2. If two variables move in the opposite direction, the covariance will be negative.
3. If there is no tendency for two variables to move one way or the other, then the covariance will be zero.

STATISTICAL MEASURES OF VARIATION

The covariance is the measure of how much two random variables move together.

1. If two variables tend to move together in the same direction, then the covariance between the two variables will be positive.
2. If two variables move in the opposite direction, the covariance will be negative.
3. If there is no tendency for two variables to move one way or the other, then the covariance will be zero.

STATISTICAL MEASURES OF VARIATION

By using the standard deviation, Pearson made it possible to measure all the points of variation on a distribution rather than the two or three points that Galton had offered in his quartile range.

STATISTICAL MEASURES OF VARIATION

The standard deviation=

$$\sqrt{\frac{sum\ of\ (\ raw\ scores - mean\ of\ observations)^2}{number\ of\ observations}}$$

FOUR PARAMETERS TO DESCRIBE THE ESSENTIAL CHARACTERISTICS OF ANY DISTRIBUTION

- **mean**: how the data cluster
- **standard deviation**: how the data spread
- **skewness**: if there were a loss of symmetry (0, negative, positive)
- **kurtosis**: if the shape of the distribution were peaked or flat (0, negative, positive)
  - A negative value = less peaked(platykurtic)
  - A positive value = more peaked(leptokurtic) (two-long-tailed kangaroos for a leptokurtic curve)
  - A zero value = a symmetrical curve(mesokurtic)

FOUR PARAMETERS TO DESCRIBE THE ESSENTIAL CHARACTERISTICS OF ANY DISTRIBUTION

- **mean**: how the data cluster
- **standard deviation**：how the data spread
- **skewness**：if there were a loss of symmetry (0, negative, positive)
- **kurtosis**：if the shape of the distribution were peaked or flat (0, negative, positive)
  - A negative value = less peaked(platykurtic)
  - A positive value = more peaked(leptokurtic) (two-long-tailed kangaroos for a leptokurtic curve)
  - A zero value = a symmetrical curve(mesokurtic)

FOUR PARAMETERS TO DESCRIBE THE ESSENTIAL CHARACTERISTICS OF ANY DISTRIBUTION

- **mean**: how the data cluster
- **standard deviation**：how the data spread
- **skewness**：if there were a loss of symmetry (0, negative, positive)
- **kurtosis**：if the shape of the distribution were peaked or flat (0, negative, positive)
  - A negative value = less peaked(platykurtic)
  - A positive value = more peaked(leptokurtic) (two-long-tailed kangaroos for a leptokurtic curve)
  - A zero value = a symmetrical curve(mesokurtic)

FOUR PARAMETERS TO DESCRIBE THE ESSENTIAL CHARACTERISTICS OF ANY DISTRIBUTION

- **mean**: how the data cluster
- **standard deviation**：how the data spread
- **skewness**：if there were a loss of symmetry (0, negative, positive)
- **kurtosis**：if the shape of the distribution were peaked or flat (0, negative, positive)
  - A negative value = less peaked(platykurtic)
  - A positive value = more peaked(leptokurtic) (two-long-tailed kangaroos for a leptokurtic curve)
  - A zero value = a symmetrical curve(mesokurtic)

```
. sum cfps2012_age, detail

                        cfps2012_age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           16            16
 5%           19            16
10%           22            16       Obs                35722
25%           30            16       Sum of Wgt.        35722

50%           44                     Mean           44.11539
                      Largest        Std. Dev.      16.87916
75%           57            93
90%           67            93       Variance       284.9061
95%           73            95       Skewness       .2293365
99%           82            99       Kurtosis       2.214244
```

Figure 4.1: 年龄的描述性统计（CFPS数据）

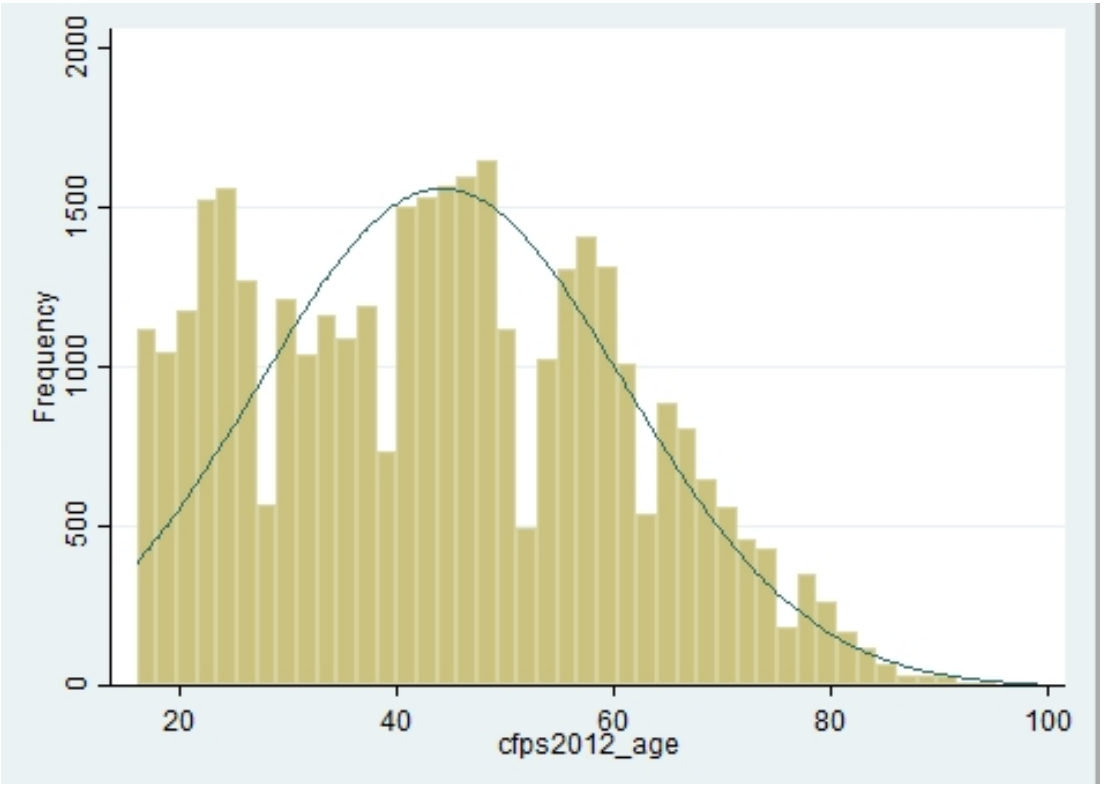


Figure 4.2: 年龄的直方图

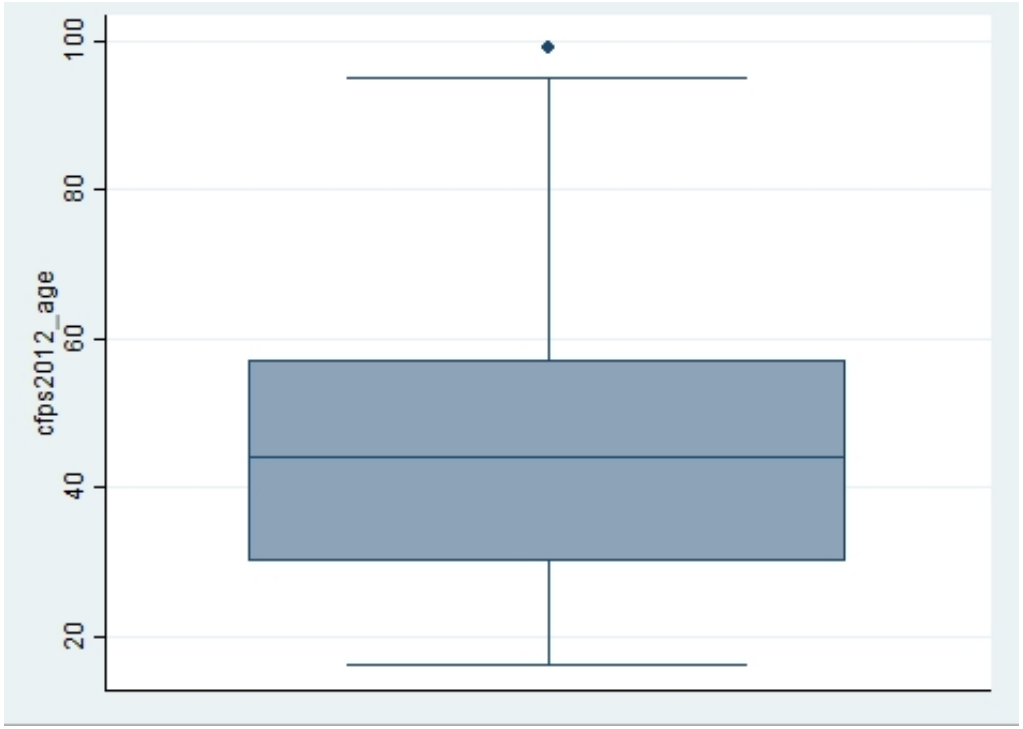


Figure 4.3: 年龄的箱线图

DESCRIPTIVE STATISTICS：两个变量

- 两个定类变量：eg.性别与志向；职业与价值观
- 两个定序变量：eg.数学成绩等级与英语成绩等级；工人积极性等级与产量等级
- 两个定距变量：eg. DV 家务劳动时间、IV 受教育年限
- 一个定类、一个定序：eg. DV 志向、IV 受教育水平
- 一个定类、一个定距：eg. IV 家庭职业背景、DV 学习成绩
- 一个定序、一个定距：eg. IV 家庭收入水平、DV 学习成绩

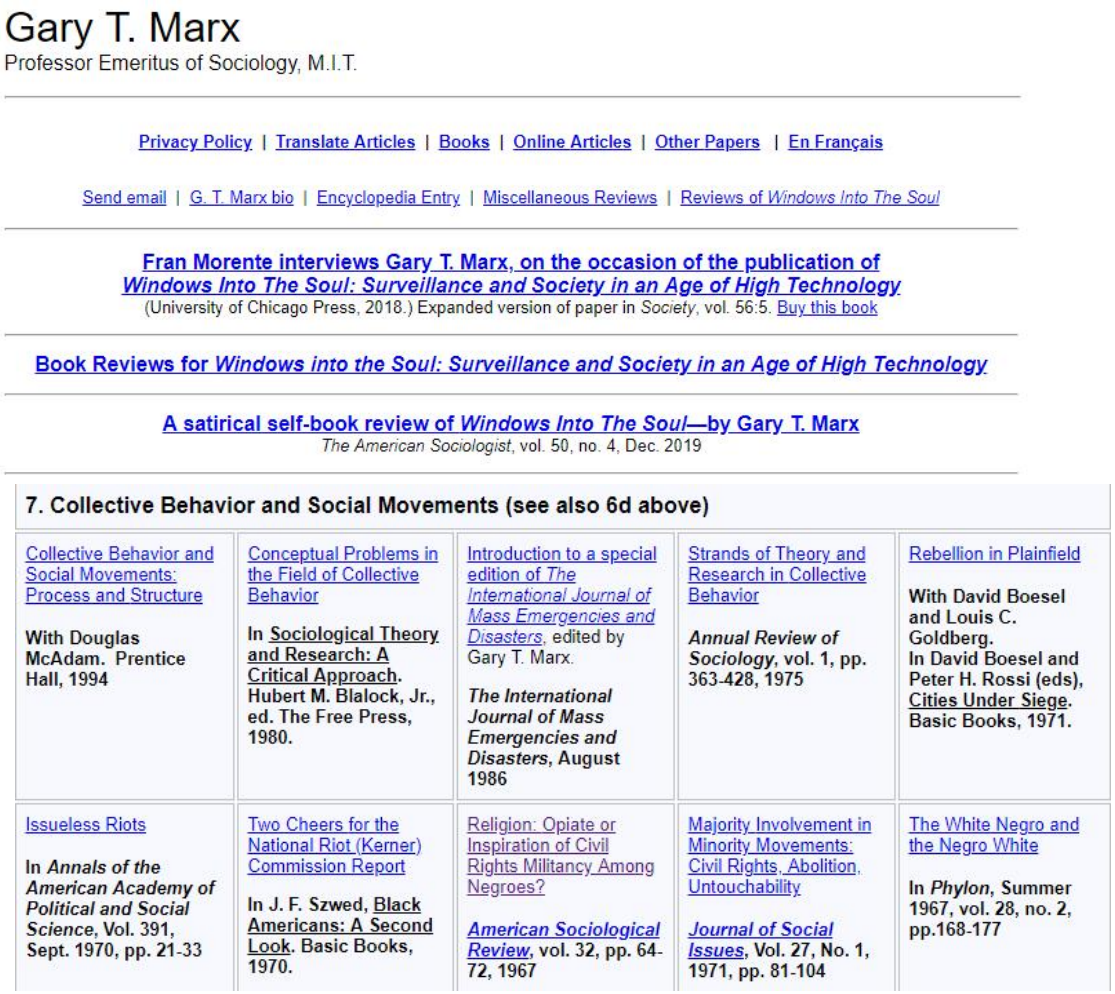
DESCRIPTIVE STATISTICS：两个变量

- 两个定类变量：Lambda, tau-y （0到1）
- 两个定序变量：Gamma, d_y, Kendall's tau, Spearman's rho （-1到1） （教材第6页）
- 两个定距变量：线性回归
- 一个定类、一个定序：Lambda, tau-y
- 一个定类、一个定距：相关比率 E^2（0到1）
- 一个定序、一个定距：相关比率 E^2（0到1）

RANK ORDER CORRELATION

Rank order correlation is the study of relationships between different rankings on the same set of items. It deals with measuring correspondence between two rankings, and assessing the statistical significance of this.

- Charles Spearman (1863-1945)
- The English statistician Maurice Kendall (1907-1983) created another ranking method of correlation in 1938, known as Kendall's tau. This method is a scheme based on the number of agreement or disagreement in ranked data.



Gary T. Marx

Professor Emeritus of Sociology, M.I.T.

Privacy Policy | Translate Articles | Books | Online Articles | Other Papers | En Français

Send email | G. T. Marx bio | Encyclopedia Entry | Miscellaneous Reviews | Reviews of *Windows Into The Soul*

Fran Morente interviews Gary T. Marx, on the occasion of the publication of *Windows Into The Soul: Surveillance and Society in an Age of High Technology*
(University of Chicago Press, 2018.) Expanded version of paper in *Society*, vol. 56:5. Buy this book

Book Reviews for *Windows into the Soul: Surveillance and Society in an Age of High Technology*

A satirical self-book review of *Windows Into The Soul*—by Gary T. Marx
The American Sociologist, vol. 50, no. 4, Dec. 2019

7. Collective Behavior and Social Movements (see also 6d above)

Collective Behavior and Social Movements: Process and Structure With Douglas McAdam. Prentice Hall, 1994	Conceptual Problems in the Field of Collective Behavior In Sociological Theory and Research: A Critical Approach. Hubert M. Blalock, Jr., ed. The Free Press, 1980.	Introduction to a special edition of *The International Journal of Mass Emergencies and Disasters*, edited by Gary T. Marx. ***The International Journal of Mass Emergencies and Disasters*, August 1986**	Strands of Theory and Research in Collective Behavior ***Annual Review of Sociology*, vol. 1, pp. 363-428, 1975**	Rebellion in Plainfield With David Boesel and Louis C. Goldberg. In David Boesel and Peter H. Rossi (eds), Cities Under Siege. Basic Books, 1971.
Issueless Riots In ***Annals of the American Academy of Political and Social Science*, Vol. 391, Sept. 1970, pp. 21-33**	Two Cheers for the National Riot (Kerner) Commission Report In J. F. Szwed, Black Americans: A Second Look. Basic Books, 1970.	Religion: Opiate or Inspiration of Civil Rights Militancy Among Negroes? ***American Sociological Review*, vol. 32, pp. 64-72, 1967**	Majority Involvement in Minority Movements: Civil Rights, Abolition, Untouchability ***Journal of Social Issues*, Vol. 27, No. 1, 1971, pp. 81-104**	The White Negro and the Negro White In ***Phylon*, Summer 1967, vol. 28, no. 2, pp.168-177**

http://web.mit.edu/gtmarx/www/garyhome.html

研究假设

1. hypothesis 1: religious people would be less militant than nonreligious people because religion gave them an other-worldly rather than this-worldly orientation, and established religious institutions have generally had a stake in the status quo and hence a conservative orientation.
2. hypothesis 2: they would be more militant because the Black churches were a major locus of civil rights militancy, and religion is an important source of universal humanistic values.
3. hypothesis 3: there would be no connection between religiosity and militancy. （教材第2页）

数据抽样

The sample consisting of a probability sample of 492 Blacks living in metropolitan areas outside the South, plus four special samples: probability samples of Blacks living in Chicago, New York, Atlanta, and Birmingham. The total number of respondents is 1119. （教材第3 页）

问卷设计

What would you say about the civil rights demonstrations over the last few years - that they have helped Negroes a great deal, help a little, hurt a little, or hurt a great deal?

1. Helped a great deal
2. Helped a little
3. Hurt a little
4. Hurt a great deal
5. Don't know

以上问题可以通过赋值（code），输入数据，形成一个变量。

Eight items were used to construct a militancy scale. Individuals were classified as militant if they gave the militant response to at least six of the eight items.

1. In your opinion, is the government in Washington pushing integration too slow, too fast, or about right?(Too slow)
2. Negroes who want to work hard can get ahead just as easily as anyone else.(Disagree)
3. Negroes should spend more time praying and less time demonstrating. (Disagree)
4. To tell the truth I would be afraid to take part in civil rights demonstrations.(Disagree)
5. Would you like to see more demonstrations or less demonstrations?(More)
6. A restaurant owner should not have to-serve Negroes if he doesn't want to.(Disagree)
7. Before Negroes are given equal rights, they have to show that they deserve them.(Disagree)
8. An owner of property should not have to sell to Negroes if he doesn't want to. (Disagree)

频数表 FREQUENCY

Cross-tabulate militancy by religiosity, that is, to count the frequency of persons with each combination of religiosity and militancy.

TABLE 1.1. Joint Frequency Distribution of Militancy by Religiosity Among Urban Negroes in the U.S., 1964.

Religiosity	Militant	Nonmilitant	Total
Very religious	61	169	**230**
Somewhat religious	160	372	**532**
Not very religious	87	108	**195**
Not at all religious	25	11	**36**
Total	**333**	**660**	**993**

比例表 PERCENT

TABLE 1.2. Percent Militant by Religiosity Among Urban Negroes in the U.S., 1964.

Militancy	Very Religious	Somewhat Religious	Not Very Religious	Not at All Religious	Total
Militant	27%	30%	45%	69%	**33%**
Nonmilitant	73	70	55	31	**67**
Total	**100%**	**100%**	**100%**	**100%**	**100%**
N	(230)	(532)	(195)	(36)	(993)

TABLE 1.3. Percentage Distribution of Religiosity by Educational Attainment, Urban Negroes in the U.S., 1964.

Religiosity	Educational Attainment		
	Grammar School	High School	College
Very religious	31%	19%	19%
Somewhat religious	57	54	45
Not very religious	12	24	25
Not at all religious	1	4	11
Total	**101%**	**101%**	**100%**
N	(353)	(504)	(136)

Source: Adapted from Marx (1967a, Table 6).

TABLE 1.4. Percent Militant by Educational Attainment, Urban Negroes in the U.S., 1964.

Militancy	Educational Attainment		
	Grammar School	High School	College
Militant	22%	36%	53%
Nonmilitant	78	64	47
Total	**100%**	**100%**	**100%**
N	(353)	(504)	(136)

Source: Adapted from Marx (1967a, Table 6).

THREE-VARIABLE CROSS-TABULATION

TABLE 1.5. Percent Militant by Religiosity and Educational Attainment, Urban Negroes in the U.S., 1964.

	Grammar School			High School			College		
Militancy	**V**	**S**	**N**	**V**	**S**	**N**	**V**	**S**	**N**
Militant	17%	22%	32%	34%	32%	47%	38%	48%	68%
Nonmilitant	83	78	68	66	68	53	62	52	32
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
N	(108)	(201)	(44)	(96)	(270)	(138)	(26)	(61)	(49)

Source: Adapted from Marx (1967a, Table 6).

*V=very religious; S=somewhat religious; N=not very religious or not at all religious.

因变量为连续变量的列联表（交叉表）分析

TABLE 3.4. Means and Standard Deviations of Income in 1979 by Education and Gender, U.S. Adults, 1980.

Education	Males	Females	Female as Percent of Male
	Means		
Post-graduate training	31,864	14,113	44
College graduate	27,227	11,789	43
Some college	19,222	13,003	68
High school graduate	16,288	10,324	63
Less than 12 years	15,855	8,399	53
Total	**20,415**	**11,135**	**55**
	Standard Deviations		
Post-graduate training	17,541	5,019	29
College graduate	14,618	6,794	46
Some college	12,912	9,704	75
High school graduate	8,935	7,573	85
Less than 12 years	11,488	6,280	55
Total	**13,790**	**7,750**	**56**

TABLE 3.5. Median Annual Income in 1979 Among Those Working Full Time in 1980, by Education and Gender, U.S. Adults (Category Frequencies Shown in Parentheses).

	Male	Female	Total
Post-graduate studies	37,500 (57)	13,750 (21)	**21,250** (78)
College graduate	23,750 (46)	11,250 (35)	**18,750** (81)
Some college	16,250 (68)	11,250 (52)	**13,750** (120)
High school graduate	16,250 (131)	9,000 (105)	**11,250** (236)
Less than 12 years	13,750 (78)	6,500 (33)	**11,250** (111)
Total	**16,250** (380)	**11,250** (246)	**13,750** (626)

Stata命令

tab policy sex, col

cross-tabulate policy (the row variable) by sex (the column variable) and compute column percentages.（教材第64页）