Development Data Boot Camp
Summary Statistics

Ge Sun

University of Notre Dame

May 23, 2023

Introduction

- You will find, almost in every empirical paper, there is a table called "Summary statistics" or "Descriptive statistics" in its "Data" section.
- It gives us a basic understanding about the key variables.
  - Think about how would we describe a school's teaching quality.
- It may have different forms, but usually it talks about two main things: **location** and **variability**

Introduction

- **Location** tells you the central value of your variable
  - mean
  - median
  - mode
- **Variability** tells you the spread of the data from the center value
  - variance
  - standard deviation
  - range (maximum, minimum)
  - other quantiles
- **Others**
  - number of observations

Stata Guidance

How can we get these numbers from Stata?

- **Location**:
  - mean sum
  - median sum,detail
  - mode mode
- **Variability**:
  - variance sum,detail
  - standard deviation sum
  - range (maximum, minimum) sum
  - other quantiles sum,detail
- **Others**
  - number of observations sum,detail

Example of descriptive statistics

TABLE 1

SELECTED SAMPLE MEANS

	CPS (1979, 1983)	PSID (1976–84)
Average log weekly earnings (1967 dollars)	4.86	...
	(.004)	
Average log hourly earnings (1967 dollars)	...	1.12
		(.004)
Highest grade completed	12.82	11.98
	(.022)	(.022)
Age − school − 5	21.21	20.87
	(.090)	(.097)
Union status	.342	.29
	(.003)	(.003)
Tenure*	8.337	83.03
	(.059)	(.659)
Nonwhite	.097	.322
	(.002)	(.003)
Ever married	.907	.854
	(.002)	(.003)
SMSA	.556	.527
	(.004)	(.004)
Unemployment rate at start of job	4.778	4.676
	(.013)	(.013)
Minimum rate since start of job	4.045	4.181
	(.014)	(.013)
1 − (emp/pop) ratio at start of job	22.34	22.49
	(.021)	(.025)
Minimum 1 − (emp/pop) ratio since start of job	22.01	22.24
	(.023)	(.025)
	Number of Observations	
1976		1,958
1977		2,130
1978		2,262
1979	9,422	2,376
1980		2,497
1981		2,433
1982		2,236
1983	9,286	2,110
1984		1,957

* Measured in years for the CPS and in months for the PSID.

* *From Beaudry and DiNardo (1991)*

Example of descriptive statistics

Table 1
Sample summary statistics

	1936–2005
Total # of person-year observations	15,883
Total # of executives	2,862
Average # of firms in each year	76
Average # of years each executive is observed	5.6
Median # of years each executive is observed	4
Fraction of obs. in firms with market value	
Ranked 1–50	39.0
Ranked 50–100	19.6
Ranked 100–200	19.1
Ranked 200–500	16.7
Ranked 500+	5.4

Based on the three highest-paid officers in the largest fifty firms in 1940, 1960, and 1990 (a total of 101 firms). Rankings by market value are based on all firms appearing in the CRSP database, which includes all publicly traded firms in the NYSE, AMEX, and NASDAQ stock exchanges. Annual market value is measured at the end of the fiscal year.

Table 2
Distribution of job titles

	Percent of observations		
	Entire sample	1936–1969	1970–2005
Chairman of the board	21.2	15.8	25.9
Vice-chairman	6.4	2.0	10.3
President	28.5	31.6	25.9
Chief executive officer	15.3	2.3	26.8
Chief financial officer	1.8	0.0	3.4
Chief operating officer	5.0	0.2	9.1
Executive or senior vice-president	21.6	15.3	27.2
Vice-president	15.2	27.8	4.1
Treasurer	1.2	2.4	0.1
Comptroller	0.6	1.3	0.1
Other job title	8.7	8.4	9.0
Director	84.7	91.7	78.6

Based on the three highest-paid officers in the largest fifty firms in 1940, 1960, and 1990 (a total of 101 firms). The sum of each column is greater than 100% because some officers hold multiple titles. Other categories not listed include "secretary," "chairman of the executive committee," and officers of subsidiaries. The row labeled "director" is the percentage of executives in the sample that are also members of the board of directors.

* *From Frydman and Saks (2010)*

Example of descriptive statistics

TABLE 1—DESCRIPTIVE STATISTICS: SCHOOL AND STUDENT CHARACTERISTICS AT BASELINE

Variable	Treatment	Control	Standardized diff
Number of schools	149	164	
Urban	0.107	0.073	0.119
	[0.311]	[0.261]	
School is co-ed	0.698	0.677	0.045
	[0.461]	[0.469]	
Males in grades 6 and 7	66.427	65.270	0.028
	[45.948]	[35.963]	
Females in grades 6 and 7	75.125	74.212	0.015
	[60.081]	[58.344]	
Number of students	7,051	7,758	
Student's age	11.833	11.854	−0.017
	[1.261]	[1.250]	
Female	0.566	0.544	0.044
	[0.496]	[0.498]	
Hindu	0.945	0.953	−0.036
	[0.227]	[0.211]	

- Standardized diff evaluates how "significant" the difference between the treatment and control group.

Why do we need a summary statistics table during the research process?

- Help us understand data, get some idea on the size of effect
  - For the same 1% increase in income level, the number is huge for billionaires but not very substantial for extremely poor people.
- Identify some peculiar values in variables, which is a nice way to know whether we have cleaned the data well.

Why do we need a summary statistics table during the research process?

- Help us understand data, get some idea on the size of effect
  - For the same 1% increase in income level, the number is huge for billionaires but not very substantial for extremely poor people.
- Identify some peculiar values in variables, which is a nice way to know whether we have cleaned the data well.
- Tips for a research assistant:
  - when you provide the summary statistics table to the advisor, the more information in the table, the better!
  - remember to adjust the unit of the tables. (1,000,000 ($) is not each to read, try 1 (million $) and specify it on the table.)

One possible format

Variable	Mean	Std	N	Min	Max	Median
age	36.9	7.07	6,550	25	50	37
wage ($)	9.57	11.41	6,550	0.30	187.5	6
edyears (years)	7.96	3.69	6,550	0	12	9
female	0.38	0.49	6,550	0	1	0