Stress-Free Stats

3) Descriptive Statistics

Jan Rovny

Sciences Po, Paris, CEE / LIEPP

Introduction

- Data and statistics
- Introduction to distributions
- Measures of central tendency
- Measures of dispersion
- Skewness

Data and Statistics

Statistics

- Descriptive statistics
  - Provide a summary of data
  - Give us an overview in which we can situate specific observations
  - Describe a sample
- Inferential statistics ($\neq$ descriptive statistics)
  - Draw inferences (generalizations) to larger populations

Data frame

- Rows are observations
  - eg: countries; individuals; country years etc.
- Columns are variables
  - Quantified characteristics of the observations

Data frame example 1

cntry	year	almp	educspend_total	Euro_atrisk	EU_empl_rate_20to64	Euro_spendRD
Austria	2000	.5	11937.2	.	71.4	1.93
Austria	2005	.6	13337.3	16.8	71.7	2.46
Austria	2010	.	16867.5	16.6	74.9	2.8
Belgium	2000	1.1	12917.7	.	65.8	1.97
Belgium	2005	1.1	17969.3	22.6	66.5	1.83
Belgium	2010	.	23395.6	20.8	67.6	2.1
Canada	2000	.4	54662.6	.	.	.
Canada	2005	.3	63658.9	.	.	.
Canada	2010	.	84166.4	.	.	.

Raw data

little overwhelming...

```
educspend_total Euro_atrisk EU_empl_rate_20to64 Euro_spendRD family_exp gdp_growth lfp_15to24 unempl_15to2
lfp_15to64 unempl_15to64 lfp_old unempl_old MARKER preprim_edspend_level
0 17.3 69 1.56 3.6 5.253 28.8322 13.6891 66.5916 4.49656 32.3952 2.12265 1 0
0 17.1 70.7 1.51 3.10143 24.7251 14.2341 68.2121 4.39723 40.5677 2.28961 1 296.87
20725.1 74.3 1.94 1.5 3.94103 70.8111 6.10399 74.3406 3.07172 38.4519 2.12766 1 1468.67
26423.2 16.7 75.1 1.9 1.7 2.04648 68.095 9.40106 75.4798 5.2912 46.8958 4.48578 1 1824.27
35085.6 15.1 76.8 1.86 1.52765 68.993 8.67052 78.2175 4.479 56.2877 3.95713 1 2413.51
7862.03 2.8 2.71945 62.7503 13.5562 75.0567 6.23276 59.6856 4.72167 1 221.577
9699.24 2.6 3.50683 62.5275 9.72636 77.315 3.86764 70.8523 1.89766 1 361
14111.3 1.81556 60.3696 17.0591 77.5447 6.71725 75.8542 3.37819 1 1038.99
97480 80.3 3 3.25358 64.6817 10.1587 80.6901 3.45572 68.0365 1.34228 1 10652
136614 16.2 78.2 1.51 2.8 2.58894 60.1782 12.0306 78.8751 4.66694 68.8279 1.7134 1 5595
174830 14.9 79.6 1.68 .478112 57.3522 9.31575 78.2474 3.6882 69.6043 1.39058 1 8493.9
35956.2 61 .64 1.2 4.2598 37.8439 35.1658 65.764 16.3703 31.3464 9.36293 1 3582.17
53743.7 45.3 58.3 .57 1.1 3.61705 33.5425 37.7724 64.6031 18.0318 32.7891 11.2341 1 5331.2
73254 27.8 64.3 .74 3.87473 34.5801 23.6667 65.3151 9.74832 36.6986 7.14597 1 7349.41
6632.44 73.5 .73 1 3.91558 45.747 8.61538 71.2235 4.15412 52.5189 3.18602 1 349.879
8044.88 26.1 72.3 .78 1.2 .775076 42.1212 16.2202 73.1983 8.05039 53.6793 6.10143 1 594.7
9721.41 25.3 70.5 1.59 1.93641 36.1289 22.7588 73.6744 11.4114 54.2928 8.89334 1 703.98
36862.1 63.5 .65 2 1.36839 45.9962 36.9612 69.8825 18.7803 24.3295 12.336 1 3901.82
57186.2 32 64.5 .51 1.9 6.65522 36.5174 29.8811 68.872 16.1954 35.0568 13.2901 1 5764.26
2781.4 20.6 64.6 .63 4.42534 30.9623 33.6335 68.6501 14.4141 45.2245 10.1394 1 260.57
68.5 1.38 1.2 4.26553 1
394512 18.5 71.1 1.44 1.1 4.00726 40.5185 15.9451 70.6668 6.6654 32.0621 4.22045 1 32581.3
2016.54 18.3 70.3 2.1 1.25845 39.9259 14.6523 71.4973 7.41074 36.478 3.96109 1 207.21
26989.8 60.7 .91 1 5.0476 48.5496 25.294 66.6966 13.9355 40.8911 9.40893 1 2331.89
38432.6 24.3 67.2 1.12 1.2 3.58365 52.3947 19.6312 71.0822 9.19242 45.9786 6.3044 1 4769.24
52091.6 26.7 62.8 1.4 -.201262 46.8711 41.4806 74.5539 19.9752 50.7497 14.2104 1 7319.47
162313 77.7 3 4.45219 52.8727 11.7308 78.97 5.87717 69.2991 6.12536 1 10642
190708 14.4 78.1 3.56 3.3 3.16078 55.5036 21.9941 80.2458 7.76657 72.8254 4.45618 1 14906
233094 15 78.1 3.39 6.55685 51.3648 24.7734 79.0454 8.74738 74.8974 5.75934 1 23716.1
```

Levels of measurement and descriptive statistics

- Different levels of measurement require different descriptive statistics
    - Nominal and ordinal measures → categorical measures
    - Interval and scale measures → continuous measures

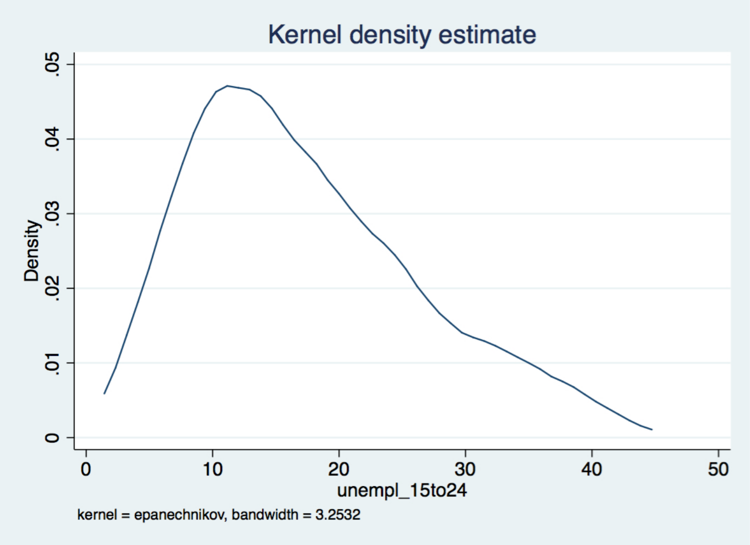
Distributions

Distribution

- Demonstrates the way in which observations are spread over possible values
- Shows the frequency of values of a sample
- To draw a distribution:
    - Collect all the values of a variable
    - Find the minimum and maximum
    - Plot all the values from the lowest to the highest

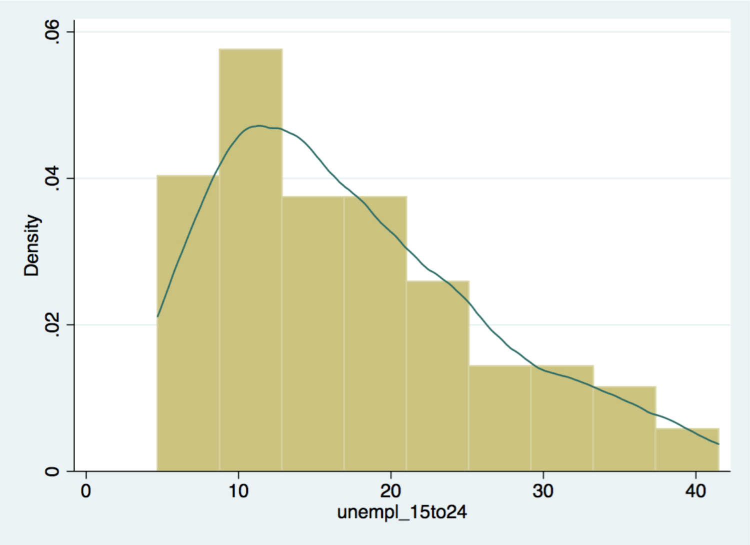
Distribution example 1

Youth unemployment rate



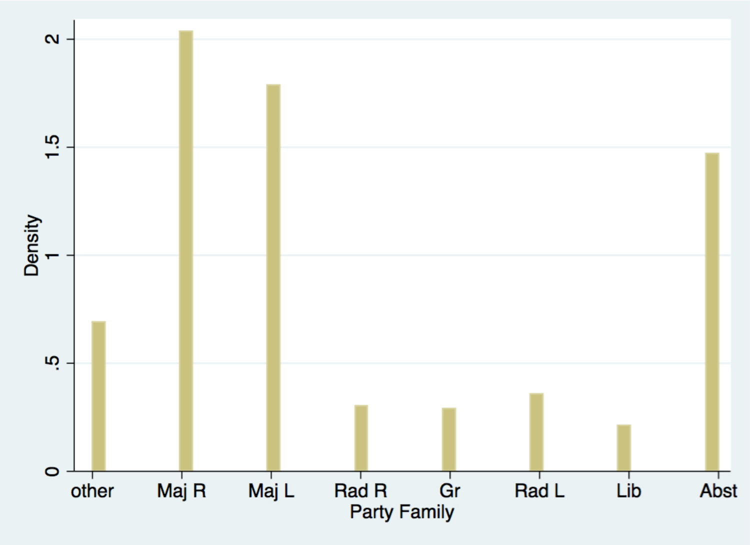
Distribution example 2

Youth unemployment rate



Distribution example 3

Voting behavior



Measures of central tendency

Measures of Central Tendency

- Measures of central tendency give different types of ‘average’ values of a variable.
- It is a summary measure of a variable.

Measure	**Calculation**	**Description**
Mode		the most frequently occurring value
Median	$\tilde{\text{X}}$	the central value separating halves of data
Mean	$\bar{X} = \mu = \frac{\sum x_i}{N}$	the arithmetic mean

Measures of Central Tendency

Different measures can be used for different levels of measurement!

Mode	nominal, ordinal, interval, scale
Median	ordinal, interval, scale
Mean	interval, scale

- Example: Identify the mode, median and mean in `(2,2,2,4,6,8,8)`

Measures of Central Tendency

Different measures can be used for different levels of measurement!

Mode	nominal, ordinal, interval, scale
Median	ordinal, interval, scale
Mean	interval, scale

- Example: Identify the mode, median and mean in `(2,2,2,4,6,8,8)`
  - Mode = 2

Measures of Central Tendency

Different measures can be used for different levels of measurement!

Mode	nominal, ordinal, interval, scale
Median	ordinal, interval, scale
Mean	interval, scale

- Example: Identify the mode, median and mean in `(2,2,2,4,6,8,8)`
  - Mode $= 2$
  - Median $= 4$

Measures of Central Tendency

Different measures can be used for different levels of measurement!	
Mode	nominal, ordinal, interval, scale
Median	ordinal, interval, scale
Mean	interval, scale

- Example: Identify the mode, median and mean in `(2,2,2,4,6,8,8)`
  - Mode $= 2$
  - Median $= 4$
  - Mean=4.571

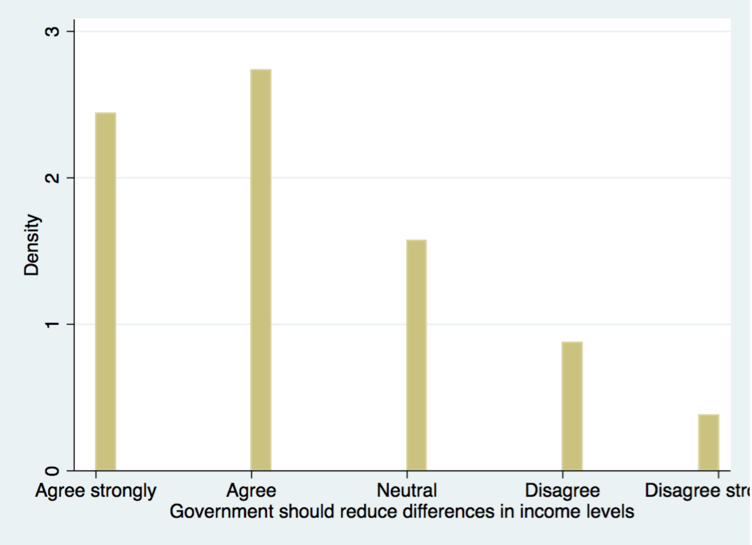
Assessing measures of central tendency

- Nominal data - histogram, frequencies

Party Family	Freq.	Percent	Cum.
other	10,614	9.68	9.68
Major right	31,234	28.48	38.15
Major left	27,452	25.03	63.18
Radical right	4,642	4.23	67.42
Green	4,449	4.06	71.47
Radical left	5,498	5.01	76.48
Minor liberal	3,238	2.95	79.44
Abstention	22,554	20.56	100.00
Total	109,681	100.00	

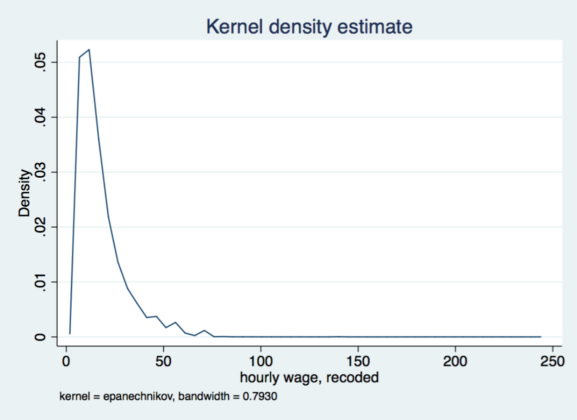
Assessing measures of central tendency

- Ordinal data - histogram, frequencies



Assessing measures of central tendency

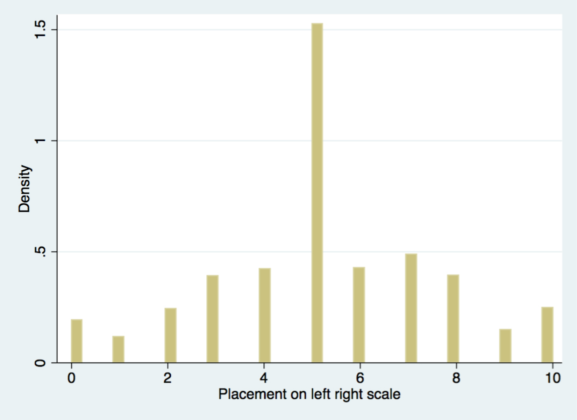
- Interval and scale data - density distribution,
- mean and standard deviation, min, max, median



Assessing measures of central tendency

Complications:

- When ordinal data is 'interval' (has equivalent unit changes along the scale), and has enough categories, we can treat it as interval data



Mean and Median in interval data

- Difference between mean and median!
- Income `(19,20,12,30,10,17,18,15,13,10)`:

Mean and Median in interval data

- Difference between mean and median!
- Income `(19,20,12,30,10,17,18,15,13,10)`:
    - $\bar{X} = 16.40$, Mode=10, $\tilde{\text{X}}$= 16.00

Mean and Median in interval data

- Difference between mean and median!
- Income (19,20,12,30,10,17,18,15,13,10):
  - $\bar{X} = 16.40$, Mode=10, $\tilde{\mathrm{X}}= 16.00$
- Enter an outlier: (19,20,12,30,10,17,18,15,13,10,575):

Mean and Median in interval data

- Difference between mean and median!
- Income (19,20,12,30,10,17,18,15,13,10):
  - $\bar{X}$ = 16.40, Mode=10, X̃= 16.00
- Enter an outlier: (19,20,12,30,10,17,18,15,13,10,575):
  - $\bar{X}$ = 67.18, Mode=10, X̃= 17.00

Mean and Median in interval data

- Difference between mean and median!
- Income `(19,20,12,30,10,17,18,15,13,10):`
  - $\bar{X} = 16.40$, Mode=10, X̃= 16.00
- Enter an outlier: `(19,20,12,30,10,17,18,15,13,10,575):`
  - $\bar{X} = 67.18$, Mode=10, X̃= 17.00
- Lesson: Mean is very sensitive to outlying data, Median much less so!

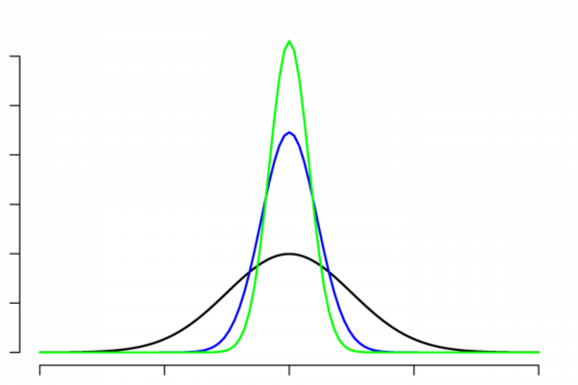
Measures of dispersion

Dispersion

- Interval data are represented by two measures
  - central tendency (mean, median)
  - dispersion
- Dispersion can be understood as spread, stretch or variability of the values

Dispersion

- Dispersion can be measured by:
    - range
    - interquartile range, 90:10 ratio
    - variance, standard deviation



Measures of Dispersion

- Tell us how close to the mean the values of the variable are.
- Is our variable ‘tightly’ around the mean, or is it widely dispersed?
- Effectively tell us how well the mean describes our variable.

Measure	**Calculation**	**Description**
Sample Variance	$s^2 = \frac{\sum(x_i - \bar{X})^2}{N-1}$	square deviation from mean
Sample Standard Dev.	$s = \sqrt{\frac{\sum(x_i - \bar{X})^2}{N-1}}$	deviation from mean

Measures of Dispersion 2

- From previous example:
- `(19,20,12,30,10,17,18,15,13,10)`:
    - $\sigma^2 = 35.82, \sigma = 5.99$
- `(19,20,12,30,10,17,18,15,13,10,575)`:
    - $\sigma^2 = 28398.96, \sigma = 168.52$
- Measures of dispersion are essential pieces of statistical information about variables!!! Mostly forgotten in mainstream media!

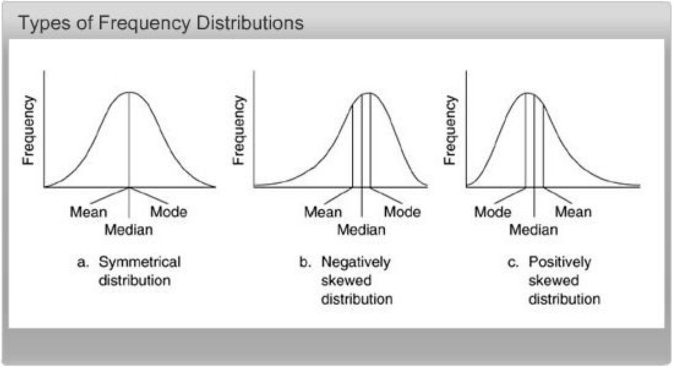
Question

- In a sample of Swedes and Brits, you notice that the highest earners are predominantly British
- Yet Swedes have higher income on average
- How is this possible?

Skewness

Skewness

- when mean=median we have a symmetrical distribution
- when mean≠median we have a skewed distribution



Types of Frequency Distributions

a. Symmetrical distribution

b. Negatively skewed distribution

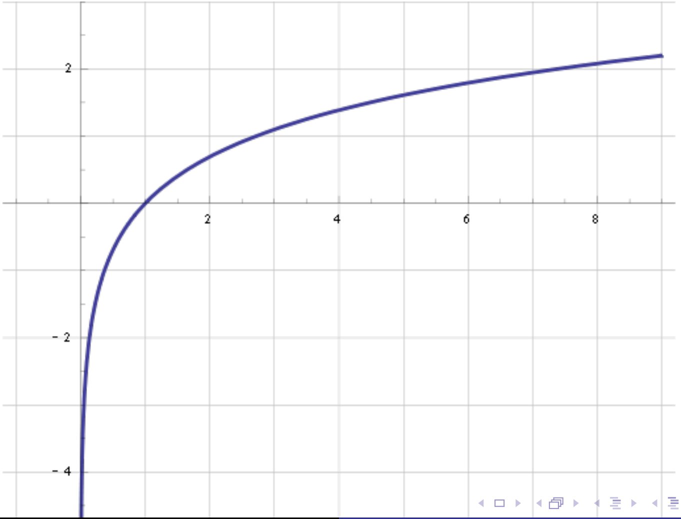
c. Positively skewed distribution

Skewness

- To deal with skew we transform variables:
  - Recode, collapsing or changing units
  - Log transformation: positive skew is fixed by logging the variable
  - Power transformation: negative skew is fixed by power transformation

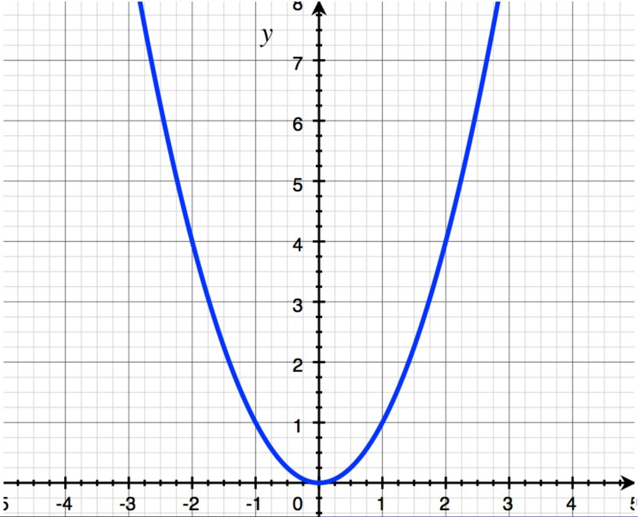
Skewness

- Why does this work?
- Log transformation "pulls" higher values in



Skewness

- Why does this work?
- Exponential transformation “pushes” higher values out


y
8
7
6
5
4
3
2
1
0
-5
-4
-3
-2
-1
1
2
3
4
5