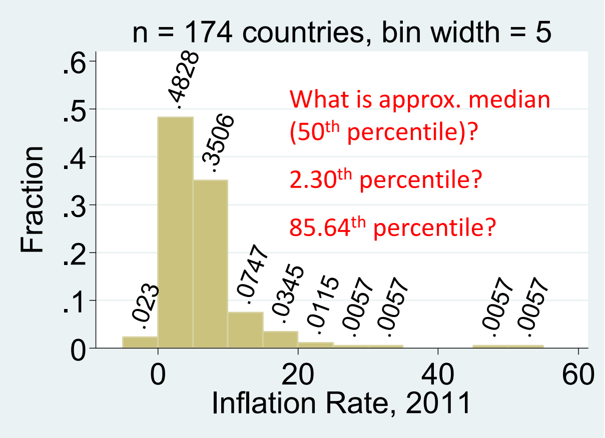
Percentiles, STATA, Box Plots, Standardizing, and Other Transformations

Lecture 3

Reading: Sections 5.7 – 5.14

Remember, when you finish a chapter make sure not to miss the last couple of boxes: "What Can Go Wrong?" and "Ethics in Action"

Measures of Relative Standing: Percentiles



What is approx. median (50th percentile)?

2.30th percentile?

85.64th percentile?

World bank data, again

Reading STATA Output

```
. su inflation_2011, detail

                       inflation_2011
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -2.517798      -4.895247
 5%     .9223603      -2.517798
10%     2.075173      -.3644478       Obs                 174
25%     3.329906      -.2833333       Sum of Wgt.         174

50%     4.977675                      Mean           6.646499
                        Largest       Std. Dev.       6.77998
75%     8.253968       26.09021
90%     12.43155       33.22422       Variance       45.96813
95%     17.71178       47.27686       Skewness       3.773002
99%     47.27686        53.2287       Kurtosis       22.85972
```

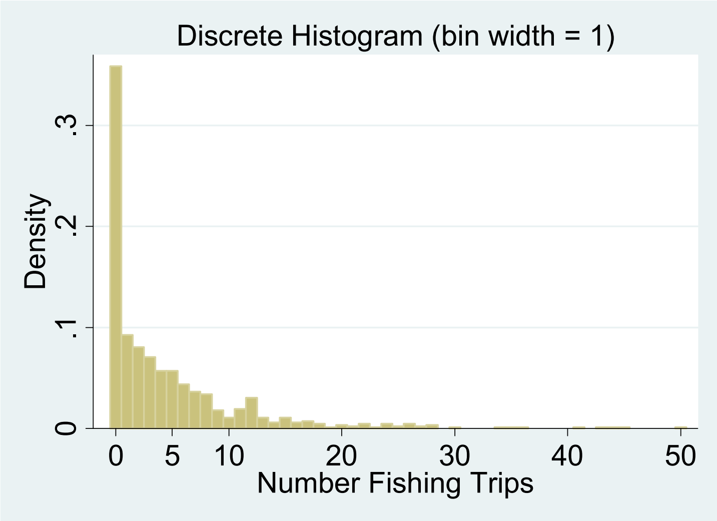
Median? Range? Sample size?

Trips	Freq.	Percent	Cum.
0	294	35.85	35.85
1	76	9.27	45.12
2	66	8.05	53.17
3	58	7.07	60.24
4	47	5.73	65.98
5	47	5.73	71.71
6	36	4.39	76.10
7	30	3.66	79.76
8	28	3.41	83.17
9	15	1.83	85.00
10	9	1.10	86.10
11	16	1.95	88.05
12	25	3.05	91.10
13	9	1.10	92.20
14	5	0.61	92.80
15	9	1.10	93.90
16	5	0.61	94.51
17	6	0.73	95.24
18	4	0.49	95.73
cont'd			

Trips	Freq.	Percent	Cum.
19	1	0.12	95.85
20	3	0.37	96.22
21	2	0.24	96.46
22	4	0.49	96.95
23	1	0.12	97.07
24	4	0.49	97.56
25	2	0.24	97.80
26	4	0.49	98.29
27	2	0.24	98.54
28	3	0.37	98.90
30	1	0.12	99.02
34	1	0.12	99.15
35	1	0.12	99.27
36	1	0.12	99.39
41	1	0.12	99.51
43	1	0.12	99.63
44	1	0.12	99.76
45	1	0.12	99.88
50	1	0.12	100.00
Total	**820**	**100.00**	

What is the median?

What is the 75th percentile?



Reading STATA Output

```
. summarize Number_of_Trips, detail;

                        Number_of_Trips
--------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                  820
25%            0              0       Sum of Wgt.          820

50%            2                      Mean             4.52439
                        Largest       Std. Dev.       6.684273
75%            6             43
90%           12             44       Variance         44.6795
95%           17             45       Skewness        2.717188
99%           30             50       Kurtosis        13.01081
```

How can the 10th percentile and the 25th percentile both be zero?

One Popular Use of Percentiles

- Quartiles:
  - 1st quartile: obs btwn 0th and 25th percentiles
  - 2nd quartile: obs btwn 25th and 50th percentiles
  - 3rd quartile: obs btwn 50th and 75th percentiles
  - 4th quartile: obs btwn 75th and 100th percentiles
- Quintiles:
  - Divide variable into fifths: e.g. top quintile includes obs btwn 80th and 100th percentiles
- Deciles:
  - Divide variable into tenths: e.g. bottom decile includes obs btwn 0th and 10th percentiles

Note: You are responsible for knowing the meaning of these terms if they appear on a test, exam, etc.

Practice Reading and Interpreting

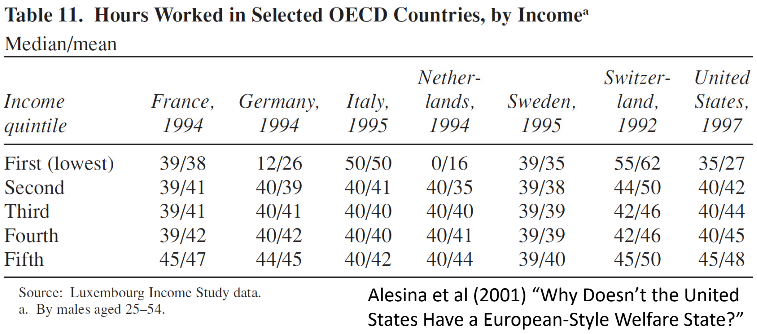


Table 11. Hours Worked in Selected OECD Countries, by Income[a]

Median/mean

Income quintile	*France, 1994*	*Germany, 1994*	*Italy, 1995*	*Netherlands, 1994*	*Sweden, 1995*	*Switzerland, 1992*	*United States, 1997*
First (lowest)	39/38	12/26	50/50	0/16	39/35	55/62	35/27
Second	39/41	40/39	40/41	40/35	39/38	44/50	40/42
Third	39/41	40/41	40/40	40/40	39/39	42/46	40/44
Fourth	39/42	40/42	40/40	40/41	39/39	42/46	40/45
Fifth	45/47	44/45	40/42	40/44	39/40	45/50	45/48

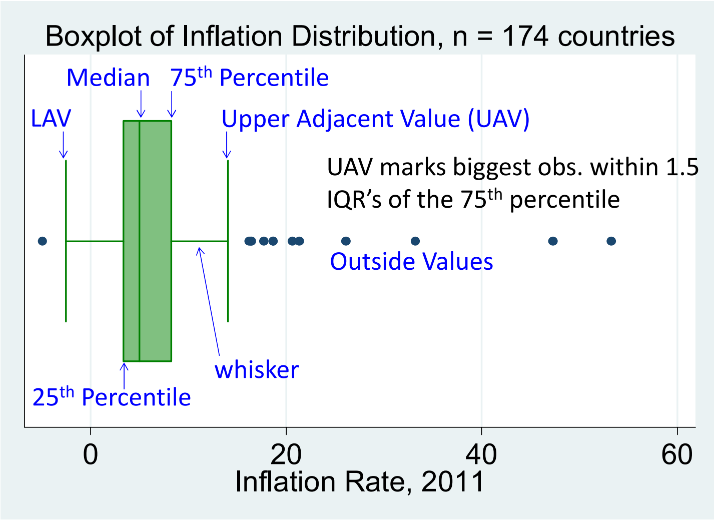
Source: Luxembourg Income Study data.
a. By males aged 25–54.

Alesina et al (2001) "Why Doesn't the United States Have a European-Style Welfare State?"

What do these numbers mean? How should they be interpreted?

Interquartile Range (IQR)

- Interquartile range: 75th percentile minus 25th percentile
  - Measures spread of middle observations
  - What does it measure?



Boxplot of Inflation Distribution, n = 174 countries

Median

75th Percentile

LAV

Upper Adjacent Value (UAV)

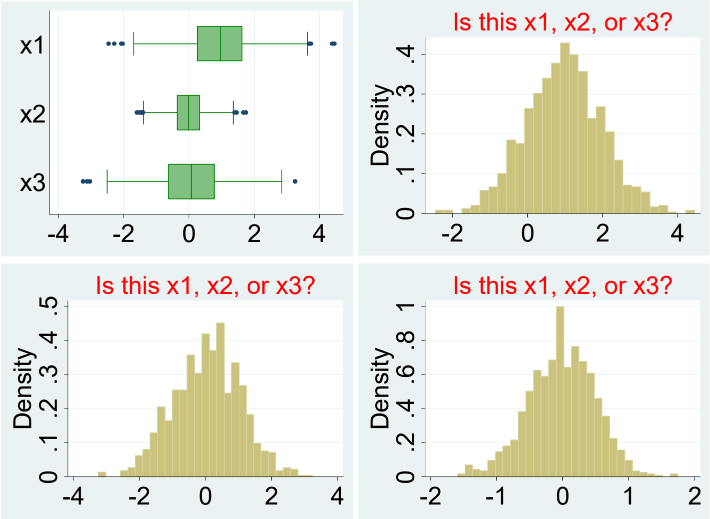
UAV marks biggest obs. within 1.5 IQR's of the 75th percentile

Outside Values

whisker

25th Percentile

10

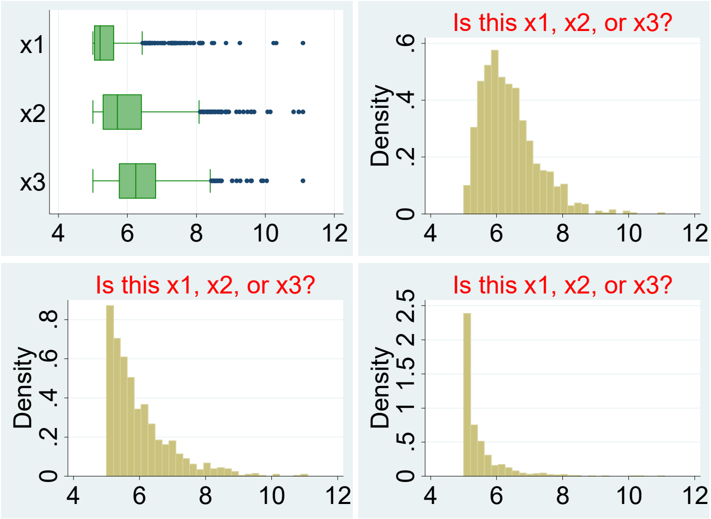


Is this x1, x2, or x3?

Is this x1, x2, or x3?

Is this x1, x2, or x3?

11

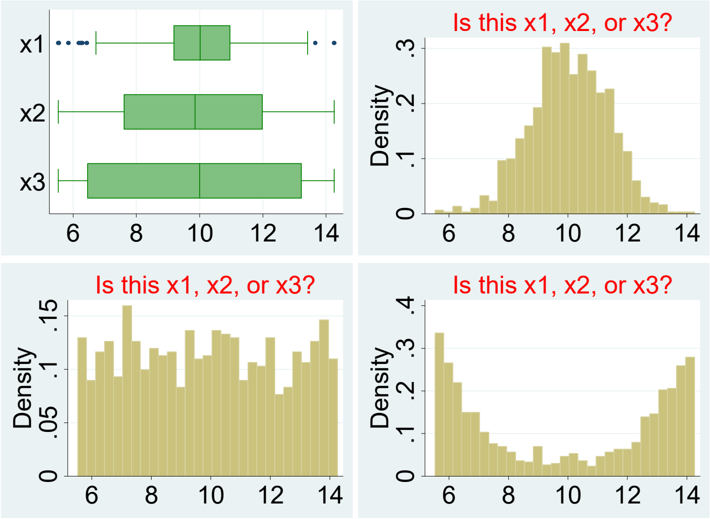


Is this x1, x2, or x3?

Is this x1, x2, or x3?

Is this x1, x2, or x3?

12



“Sunlight and Protection Against Influenza”

Table 1: Summary Statistics

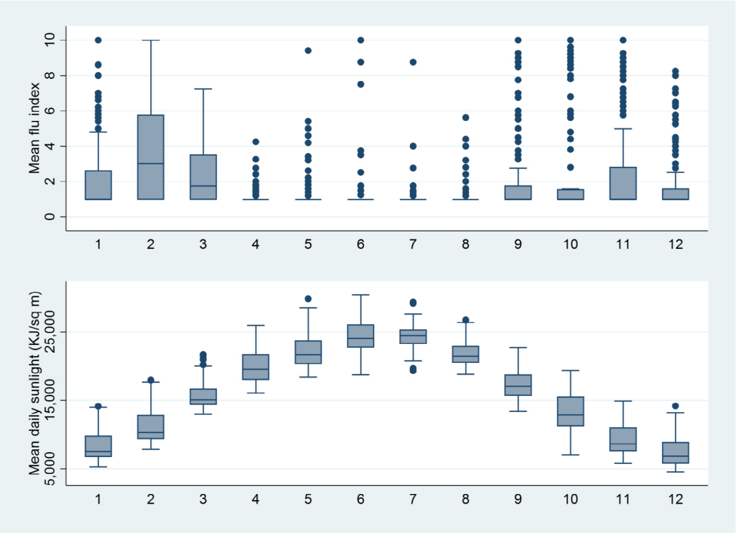
	(1) N	(2) Mean	(3) StDev	(4) Min	(5) Max
Flu index	1,404	2.000	2.139	1	10
Sunlight (kJ/m^2/day)	1,404	15,771	6,509	4,576	30,334
Population Density (individuals/mi^2)	1,404	197.2	269.5	5.8	1,195
Temperature (°F)	1,404	54.0	17.9	5.1	94.3
Days/month temp <15°F	1,404	2.0	4.7	0	29.8
Specific humidity (g water vapor / kg air)	1,404	10.8	6.4	1.8	29.7
Days/month specific humidity < 6 g/kg	1,404	9.8	10.5	0	31

Note: Unit of observation is a year-month for each of the 36 contiguous [U.S.] states that have complete flu and sunlight data.

Which kind of data are these: cross-sectional, time series, or panel?

Why 1,404 observations? These are monthly data from Oct. 2008 to Dec. 2011 (39 months) for 36 states (39*36=1,404).

Slusky and Zeckhauser (2018), http://www.nber.org/papers/w24340.pdf



Jan is 1, Feb is 2, ... Each month has 108 obs (36 states*3yrs) except Oct, Nov, and Dec have 144 obs (36 states*4yrs). N = 1,404 (=9*108 + 3*144)

Outliers

- Outliers: extremely large or small values different from the bulk of the data
- Robust: not sensitive to outliers
  - Is the sample mean a robust measure of central tendency?
    - Is the sample median robust?
    - However, the mean retains more information from sample & has useful statistical properties
  - Is the IQR robust? variance?

Charitable Donors: Stats Can

http://www5.statcan.gc.ca/cansim/a05?lang=eng&id=1110002&pattern=1110002&searchTypeByValue=1&p2=35

Donors and donations	2011
Number of taxfilers[4]	24,841,630
Number of donors[2,3]	5,709,700
Percentage of donors aged 0 to 24 years[2,3,6]	3
Percentage of donors aged 25 to 34 years[2,3,6]	12
Percentage of donors aged 35 to 44 years[2,3,6]	17
Percentage of donors aged 45 to 54 years[2,3,6]	23
Percentage of donors aged 55 to 64 years[2,3,6]	21
Percentage of donors aged 65 years and over[2,3,6]	25

[2] Charitable donor is defined as a taxfiler reporting a charitable donation amount on line 340 of the personal income tax form.

Average Age of Donors?

Section 5.7 "Grouped Data" tells how to *approximate* the mean & s.d. with grouped data

% aged 0 to 24	3
% aged 25 to 34	12
% aged 35 to 44	17
% aged 45 to 54	23
% aged 55 to 64	21
% aged 65 and over	25

$$Mean \approx 0.03 \ast 21 + 0.12 \ast 29.5 + 0.17 \ast 39.5 + 0.23 \ast 49.5 + 0.21 \ast 59.5 + 0.25 \ast 70 \approx 52.3 \text{ years}$$

What if we use 75 years old for last category? Then mean ≈ 53.5.

What if we use 12 years old for first category? Then mean ≈ 52.0.

Logic of Calculation: Smaller Example

- Survey a random sample of 40 A&S students asking how many courses are you currently taking. A tabulation:

```
num_courses |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |          3        7.50        7.50
          4 |          7       17.50       25.00
          5 |         28       70.00       95.00
          6 |          2        5.00      100.00
------------+-----------------------------------
      Total |         40      100.00
```

$$\bar{X} = \frac{\sum_{i=1}^{40} x_i}{n} = \frac{\sum_{i=1}^{3} 2 + \sum_{i=1}^{7} 4 + \sum_{i=1}^{28} 5 + \sum_{i=1}^{2} 6}{40} = \frac{3 * 2 + 7 * 4 + 28 * 5 + 2 * 6}{40}$$

$$= 0.075 * 2 + 0.175 * 4 + 0.7 * 5 + 0.05 * 6 = 4.65$$

Similarly for standard deviation

```
num_courses |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |          3        7.50        7.50
          4 |          7       17.50       25.00
          5 |         28       70.00       95.00
          6 |          2        5.00      100.00
------------+-----------------------------------
      Total |         40      100.00
```

$$s = \sqrt{\frac{\sum_{i=1}^{40}(x_i - \bar{X})^2}{n-1}}$$

$$= \sqrt{\frac{\sum_{i=1}^{3}(2-4.65)^2 + \sum_{i=1}^{7}(4-4.65)^2 + \sum_{i=1}^{28}(5-4.65)^2 + \sum_{i=1}^{2}(6-4.65)^2}{40} * \frac{40}{39}}$$

$$= \sqrt{(0.075(2-4.65)^2 + 0.175(4-4.65)^2 + 0.7(5-4.65)^2 + 0.05(6-4.65)^2)\frac{40}{39}}$$

$= 0.89$ And, if you ignore 40/39, you get 0.88 (very close to right answer)

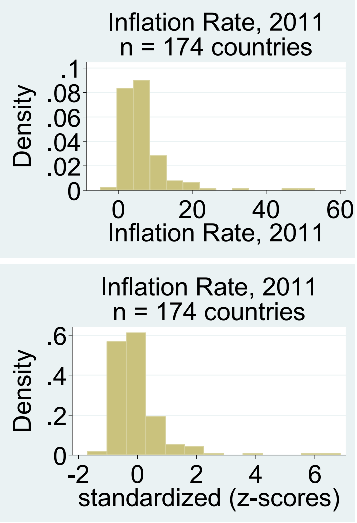
Standard Deviation of Age of Donors?

% aged 0 - 24 [21]	3
% aged 25 - 34 [29.5]	12
% aged 35 - 44 [39.5]	17
% aged 45 - 54 [49.5]	23
% aged 55 - 64 [59.5]	21
% aged 65 & over [70]	25

$$\begin{aligned} & s^2 \\ & \approx 0.03(21 - 52.3)^2 \\ & + 0.12(29.5 - 52.3)^2 \\ & + 0.17(39.5 - 52.3)^2 \\ & + 0.23(49.5 - 52.3)^2 \\ & + 0.21(59.5 - 52.3)^2 \\ & + 0.25(70 - 52.3)^2 \\ & = 210.6 \text{ years}^2 \\ & s.d. \approx \sqrt{210.6} = 14.5 \text{ years} \end{aligned}$$

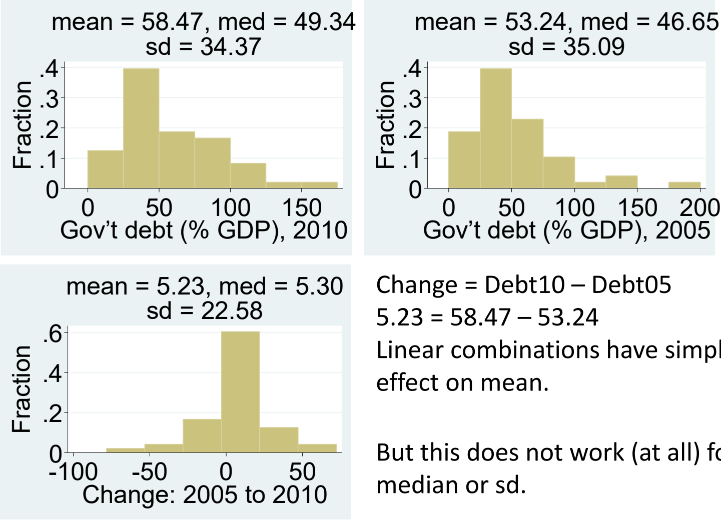
Standardization (“z-scores”)

- Standardize: $z = \frac{x - \bar{X}}{s_x}$
  - z: how many s.d.’s a value is from the mean (+ if above; - if below)
  - Z has a mean of 0 and s.d. of 1 and no units
  - Eg: mean inflation 6.64, s.d. 6.78; 2.91 in Canada: z=-0.55=(2.91-6.64)/6.78
  - What does -0.55 mean?



Linear Transformations

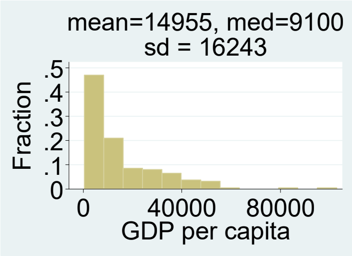
- Linear transformation can be written as Y = a + bX where a and b are constants
  - Linear transformation of X?
    - Y = 200 – X
    - Y = X^2 – 1 = (X – 1)(X + 1)
    - Y = (X - 10)/2
  - Linear transformations change scale of a variable but not shape of the distribution
  - Standardization is a linear transformation

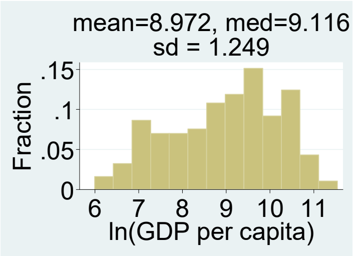


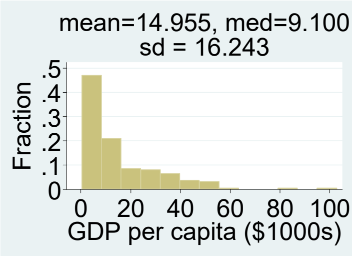
Change = Debt10 – Debt05
5.23 = 58.47 – 53.24
Linear combinations have simple effect on mean.

But this does not work (at all) for median or sd.

World Bank data again, Central gov’t debt, n = 48 countries







Non-linear transformations (natural log is very popular) can often transform skewed data to be more symmetric.

Linear transformations (such as changing units) do not affect the shape at all.

CIA data again, US$, PPP, 2012 est., n = 185 countries