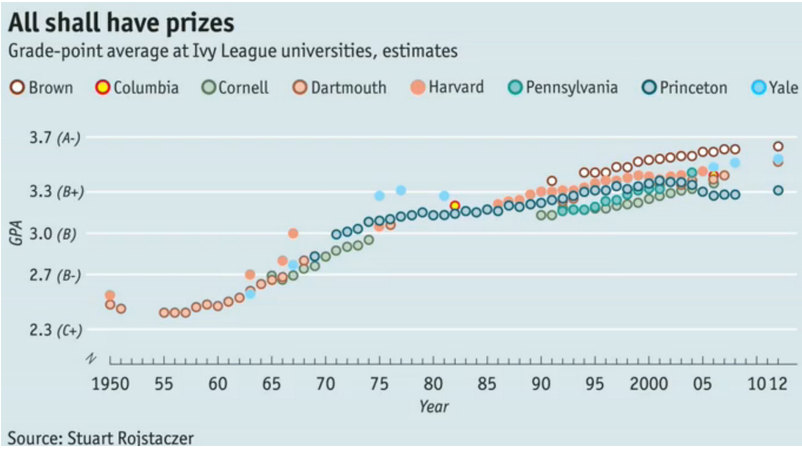


The Economist, September 6, 2014

1

Histograms, Central Tendency, and Variability

Lecture 2

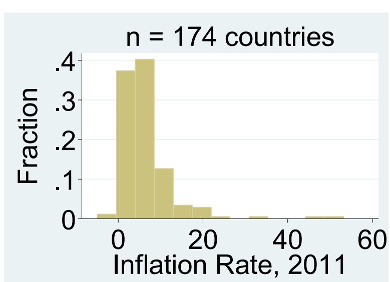
Reading: Sections 5.1 – 5.6

Includes ALL margin notes and boxes: "For Example," "Guided Example," "Notation Alert," "Just Checking," "Optional Math Boxes", "What Can Go Wrong?" and "Ethics in Action"

2

Histogram

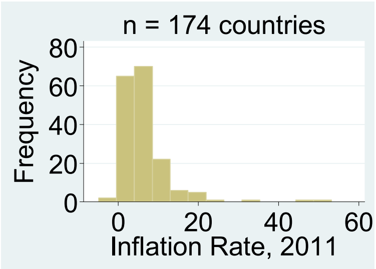
- Histogram graphically describes how a single variable containing interval data is distributed
- Range of data divided into non-overlapping and equal width classes (bins) that cover range of values

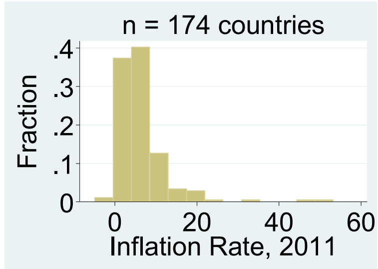


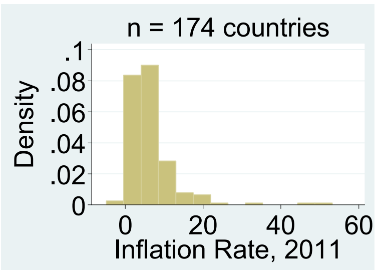
How many bins? Width of bins?

http://data.worldbank.org/indicator/FP.CPI.TOTL.ZG

3





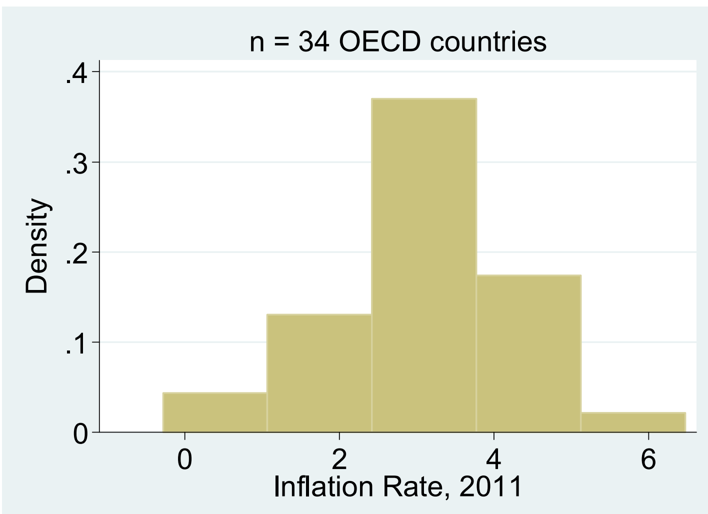


Frequency histogram: Bar height number of observations in bin

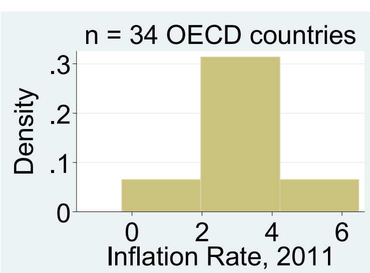
Relative frequency histogram: Bar height fraction of obs. in bin

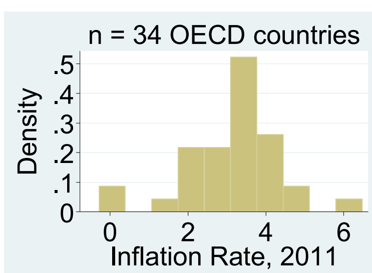
Density histogram: Bar area measures the fraction of observations in bin

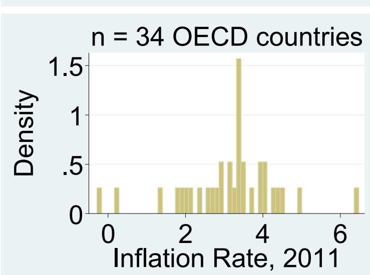
4



5







Number of bins changes the appearance of the histogram

Sturges' formula: # of bins = 1 + 3.3*log(n) [Note log base 10.]

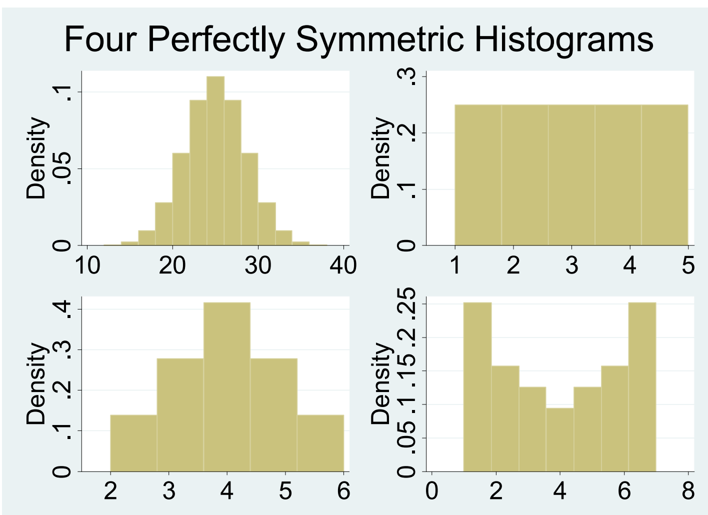
OECD inflation: 1 + 3.3*log(34) = 6.05 ≈ 6 [but STATA picked 5]

6

Shape of Things

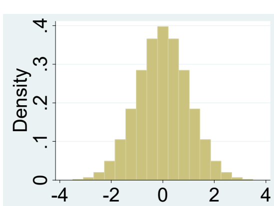
- Histogram gives overview of a variable with a single picture
  - Can make informal inferences about the shape of population
- Symmetric: If draw an imaginary line at center, have mirror image on each side
- Bell/Normal/Gaussian
- Positively skewed: long tail to right (aka right skewed)
- Negatively skewed: long tail to left (aka left skewed)
- Modality: # major peaks

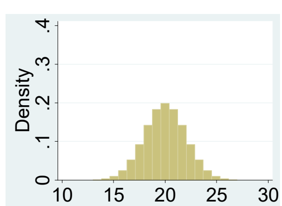
7



8

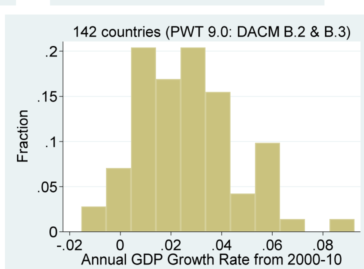
Two Perfectly Bell Shaped Histograms



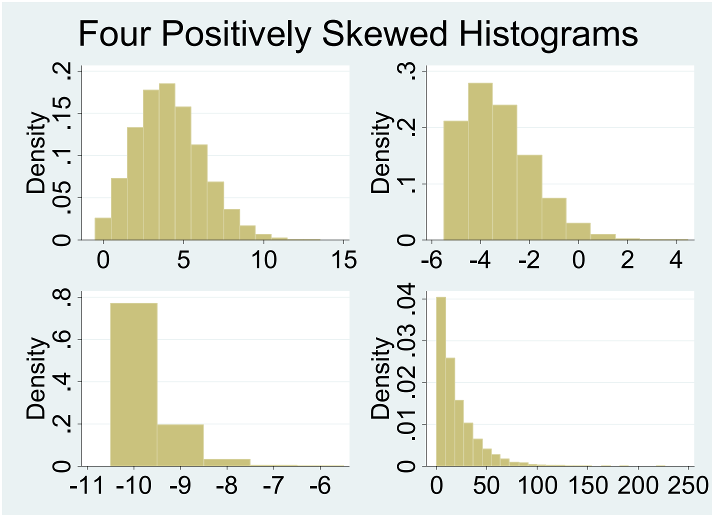


But histograms of real data will never be *perfect:* we always mean *approximately*

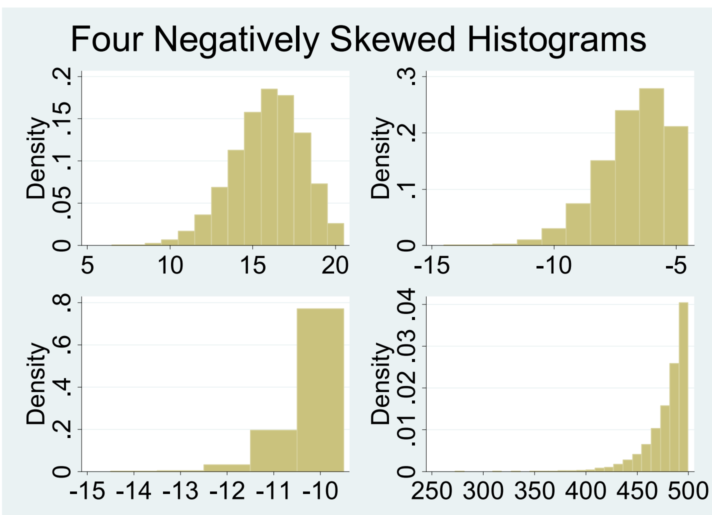
For example, we'd describe the histogram to the right as Normal (Bell) shaped



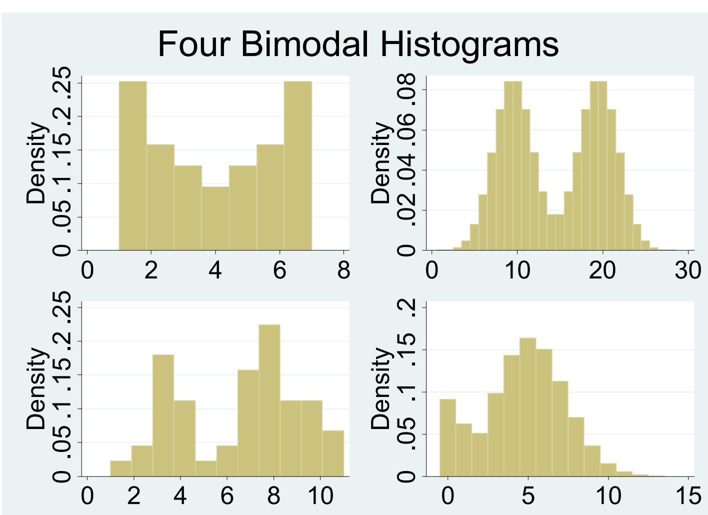
9


Four Positively Skewed Histograms
Density
Density
Density
Density

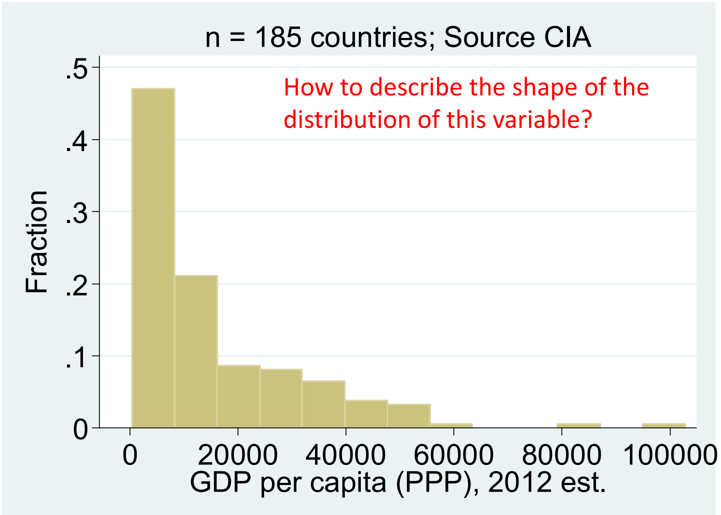
10


Four Negatively Skewed Histograms
Density
Density
Density
Density

11

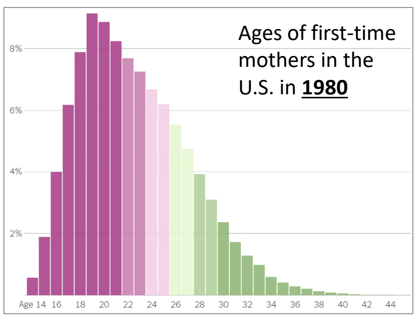

Four Bimodal Histograms
Density
Density
Density
Density

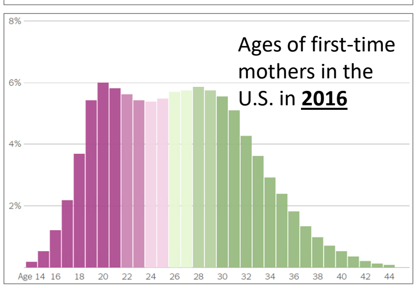
12



https://www.cia.gov/library/publications/the-world-factbook/rankorder/rawdata_2004.txt

13





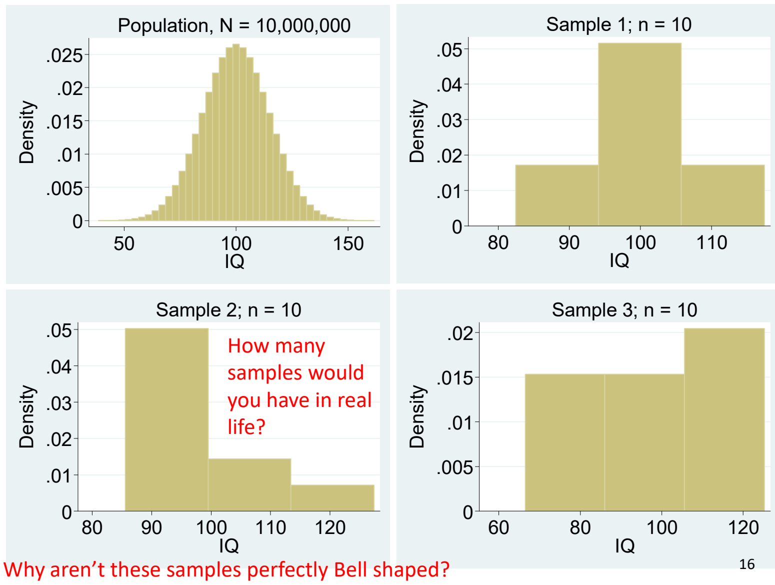
The New York Times, August 4, 2018, “The Age That Women Have Babies: How a Gap Divides America”

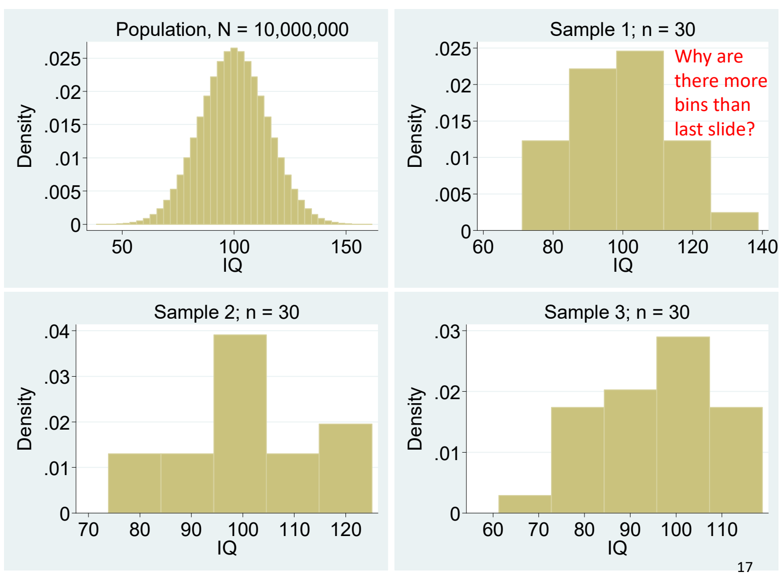
14

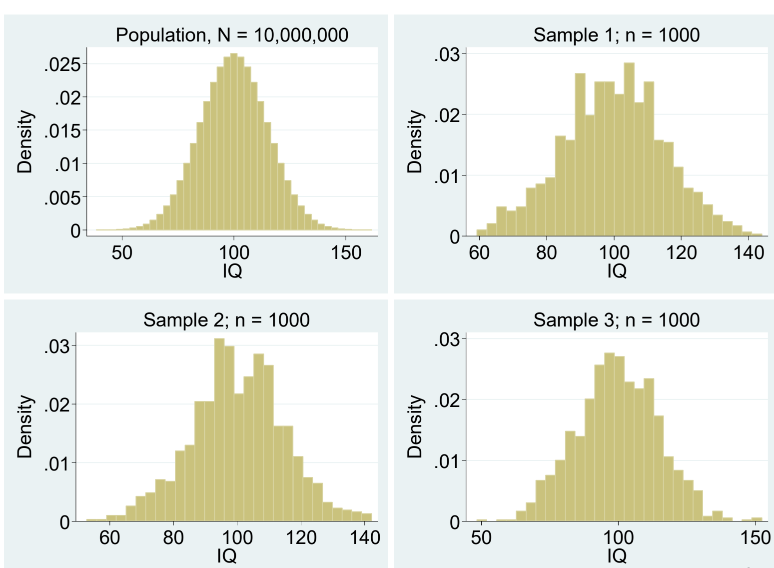
Samples vs. Populations

- Sample is a random subset of population
  - Sampling noise: Chance differences between population and a random sample
    - Driven by the sample size, not sample size relative to the population size, which is assumed infinite (pp. 30 – 31, “The Sample Size is What Matters”)
  - Informal inference: consider sample size (n)
    - Never see the perfect forms (Plato): statements about shape always approximate
    - “Nearly Normal Condition”

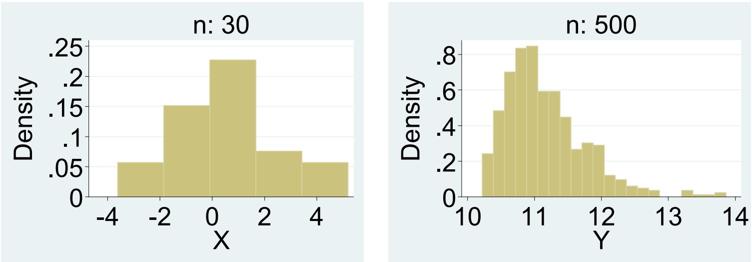
15


Population, N = 10,000,000
Sample 1; n = 10
Sample 2; n = 10
How many samples would you have in real life?
Sample 3; n = 10
Density
IQ
Why aren't these samples perfectly Bell shaped?
16


Population, N = 10,000,000
Sample 1; n = 30
Why are there more bins than last slide?
Sample 2; n = 30
Sample 3; n = 30
Density
IQ
17


Population, N = 10,000,000
Sample 1; n = 1000
Sample 2; n = 1000
Sample 3; n = 1000
Density
IQ
18

What to Conclude About Shape?

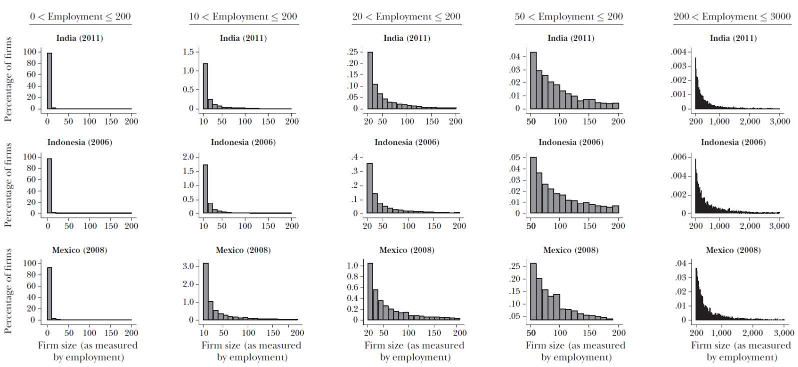


Is the graph on the left symmetric? Bell shaped?

Is the graph on the right symmetric? Bell shaped? Bi-modal?

19

Distribution of Firm Size as Measured by Number of Workers

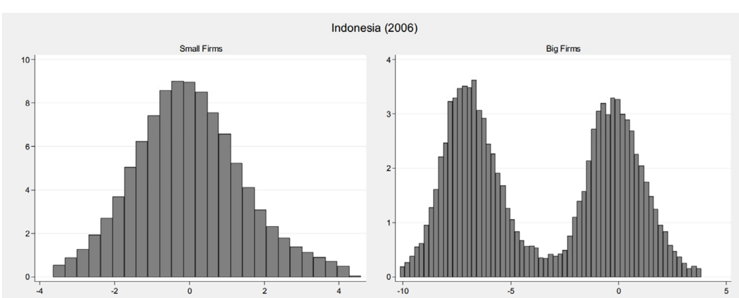


Source: We use microdata from the manufacturing sector in the Mexican Economic Census, the Indonesian Economic Census, and India's Annual Survey of Industries and National Sample Survey (Schedule 2). See footnote 1.
Notes: The figure shows distribution of firm size measured by the number of workers. The bin size is 10 workers, and each bin contains the upper bound and not the lower bound. For all graphs, the y-axis indicates the share of all firms in the specified size. The different columns truncate the x-axis in different ways to focus on different parts of the distribution.

Hsieh and Olken (2014) JEP "The Missing 'Missing Middle'" Summer 2014 http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.28.3.89

20

Figure A1: Indonesia's Distribution of Value-Added per Capita



"There is a clear bimodality in the distribution of value-added/capital for the large firms. However, the capital questionnaire for large firms was ambiguous as to whether the results were to be entered in thousands or millions of Rupiah. Our best guess is that approximately half the firms used thousands and half used millions." http://www.aeaweb.org/jep/app/2803/28030089_app.pdf

So if the real distribution of value-added/capital for large firms is Normal, which explains the bimodal shape: sampling error or non-sampling error?

21

Summary Statistics

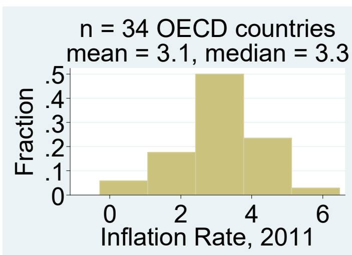
- Statistics (i.e. summary statistics) give a concise idea of what data "look like"
  - For a single variable, statistics can give numeric measures of:
    - Central tendency: mean and median
    - Variability: range, variance, standard deviation, coefficient of variation, IQR
    - Relative standing: percentiles
  - For two variables, also measure relationship

22

Mean and Median

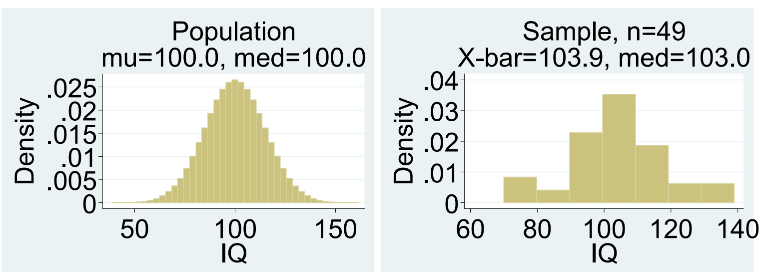
- Population mean, a parameter: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
- Sample mean, a statistic: $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Which is subject to sampling error?

- Median is the middle obs. after sorting
  - if even # of obs., average 2 middle ones



23

Normal Distribution



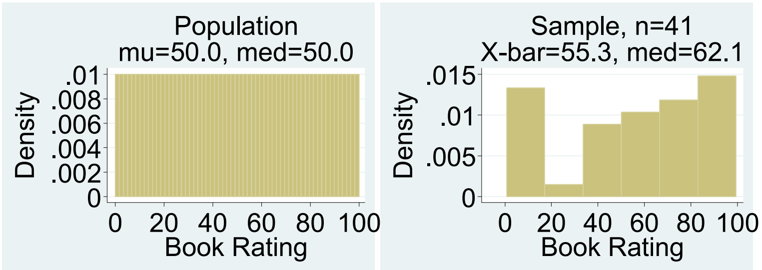
Is $\bar{X}$ (a statistic) equal to μ (a parameter)?

Why is the population mean equal to the population median?

Why isn't the sample mean equal to the sample median?

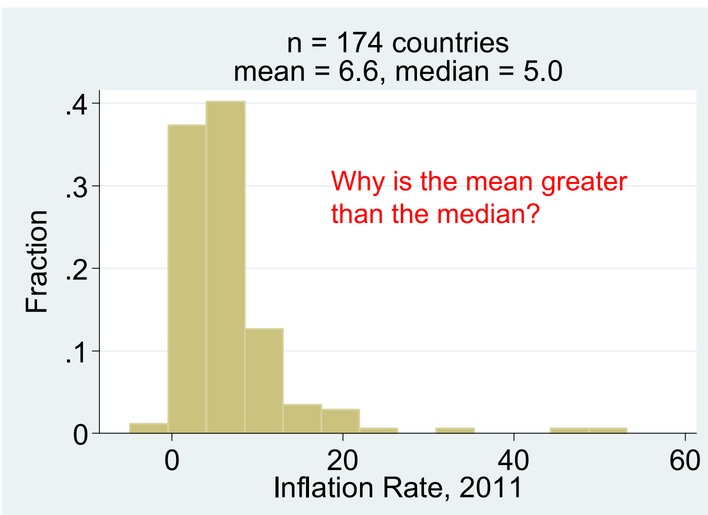
24

A Symmetric Distribution (Uniform)



Why is the population mean equal to the population median?

Why is the sample median different from the population median?



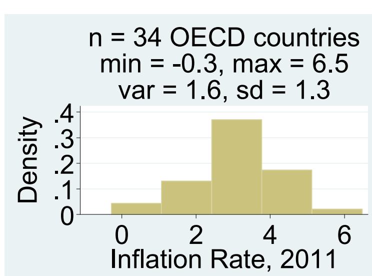
Measures of Variability (Spread)

- Range: max – min
- Variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

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

- Standard deviation: $s = \sqrt{variance}$
- Coefficient of variation (textbook)



Breaking Down Variance

- Numerator: "total sum of squares" (TSS)
  - If all sampled countries have 3% inflation (x_i = 3 for all i), what would TSS & s² be?
- Denominator: ν ("nu")
  - Only n – 1 free obs left after calculate mean
- Units of variance?
  - How about s.d.?

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

$$TSS = \sum_{i=1}^{n}(x_i - \bar{X})^2$$

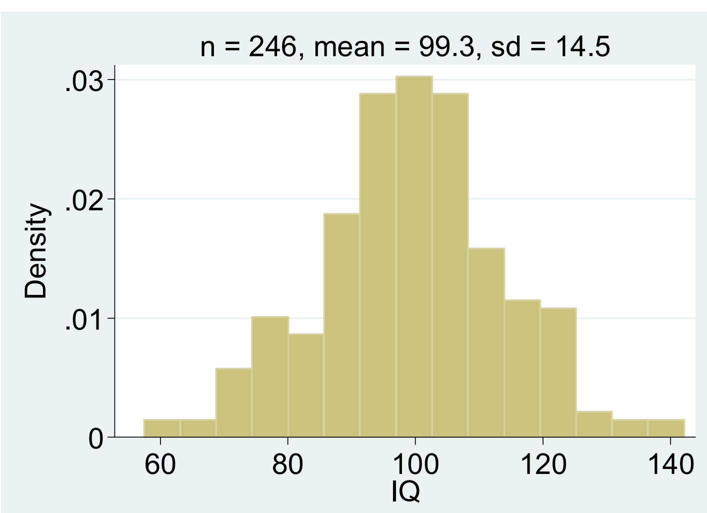
$$degrees\ of\ freedom: \\ \nu = n - 1$$

28

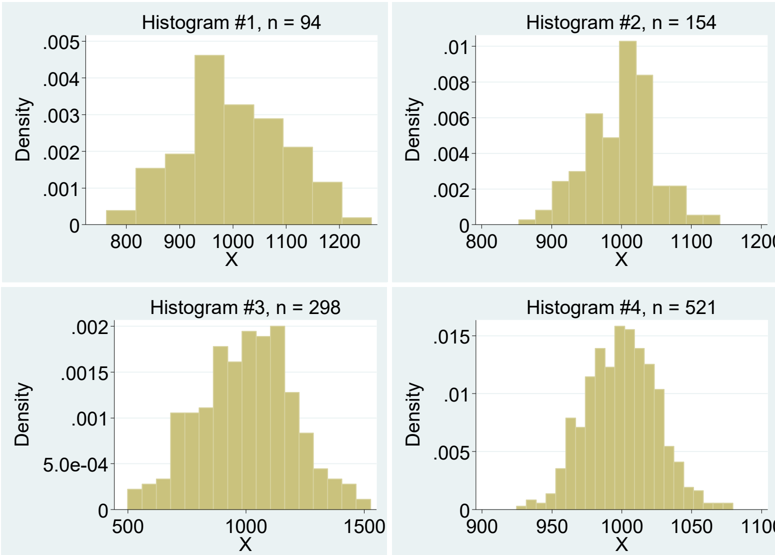
Empirical Rule (Normal/Bell)

- If a random sample is drawn from a Normal population then about:
  - 68.3% of observations will lie within 1 s.d. of the mean (i.e. between $\bar{X} - s$ and $\bar{X} + s$)
  - 95.4% of observations will lie within 2 s.d. of the mean (i.e. between $\bar{X} - 2s$ and $\bar{X} + 2s$)
  - 99.7% of observations will lie within 3 s.d. of the mean (i.e. between $\bar{X} - 3s$ and $\bar{X} + 3s$)
- "Empirical Rule" only applies if Normal

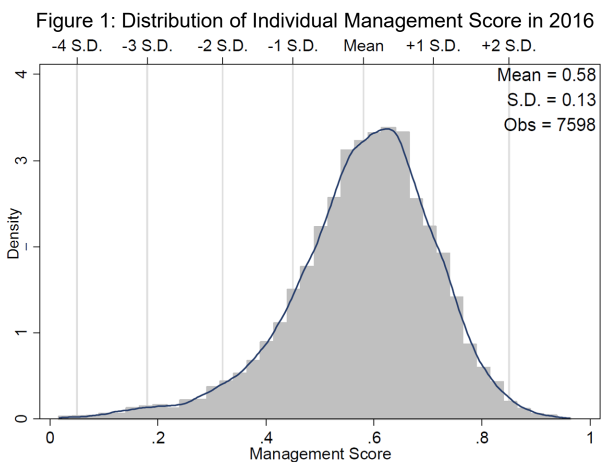
29



30



Approximately, what is the s.d. of the variable in each histogram? 31



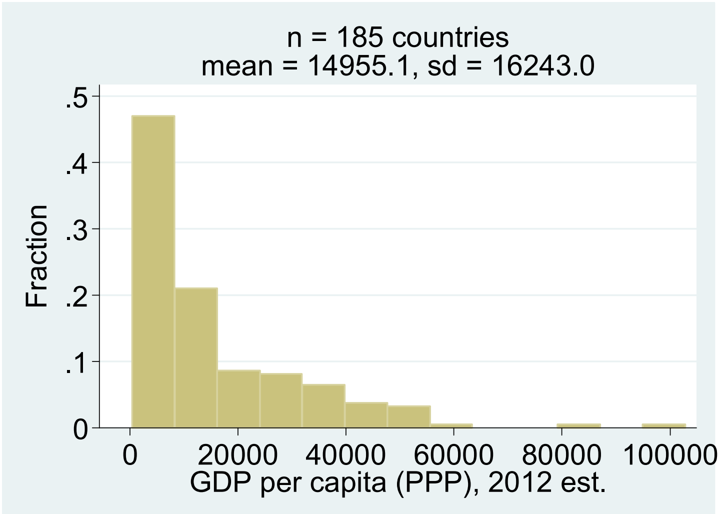
Notes: The management score is unweighted average of the score for each of the 16 questions, where each question is first normalized to be on a 0-1 scale. The sample is all 2016 CEES surveyors with at least 11 non-missing responses to management questions and [select firms].

"Do CEOs Know Best? Evidence from China" (2018) http://www.nber.org/papers/w24760 32

Chebysheff's Theorem

- At least 100*(1–1/k^2)% of observations lie within k s.d.'s of the mean for k>1
  - At least 75% of obs. lie within 2 s.d. of mean
    - 1 – 1/2^2 = 3/4
  - At least 89% of obs. lie within 3 s.d. of mean
    - 1 - 1/3^2 = 8/9
  - Can be applied to all samples no matter how population is distributed
  - What about within one s.d.?

33


n = 185 countries
mean = 14955.1, sd = 16243.0
Fraction
.5
.4
.3
.2
.1
0
0
20000
40000
60000
80000
100000
GDP per capita (PPP), 2012 est.

34