Analysis of Economics Data
Chapter 2: Univariate Data Summary

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 2: Univariate Data Summary

- Univariate data are a single series of data on a single variable.
  - e.g. annual earnings, individual income, ...
- Summarize data features using
  - summary statistics
    - ★ mean, median, standard deviation, ...
  - graphical methods
    - ★ box plot, histograms, smoothed histograms (kernel density estimates), line charts, bar charts, ..

Outline

- Datasets: EARNINGS, REALGDPPC, HEALTHCATEGORIES, FISHING, MONTHLYHOMESALES.

2.1 Summary Statistics for Numerical Data

- Observations for a **sample** of size n are denoted

$$x_1, x_2, ..., x_n.$$

- Notation: x_1 is the first observation, ... x_n is the n^{th} observation
  - cross-section data: typical observation is the i^{th}, denoted x_i
  - time series data: more customary to use the subscript t.
- Example: Sample mean or average

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n} \sum\nolimits_{i=1}^{n} x_i.$$

Summary Statistics Example: Earnings

- Summary statistics, rounded to the nearest dollar, for the earnings data on full-time working women aged 30 in 2010.

Statistic	Value
Mean	41,413
Standard deviation	25,527
Minimum	1,050
Maximum	172,000
Number of Observations	171
Variance	651,630,282
Upper quartile (75th percentile)	50,000
Median (50th percentile)	36,000
Lower quartile	25,000
Skewness	1.71
Kurtosis	7.32

Central Tendency

- **Mean:** is the average

$$\bar{x} = (x_1 + x_2 + \ldots + x_n)/n = \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i.$$

  - e.g. Sample $\{8, 3, 7, 6\}$ then $\bar{x} = (8+3+7+6)/4 = 6$
- **Median**: mid-point of the ordered observations
  - e.g. Sample $\{8, 3, 7, 6\}$ when ordered is $\{3, 6, 7, 8\}$
  - median is average of the middle two values $= (6+7)/2 = 6.5$.
- Other measures
  - **Mid-range**: average of the smallest and largest values
  - **Mode**: most common value (not useful for continuous data).
- Most often the mean is used.

Data Dispersion or Spread: Standard Deviation

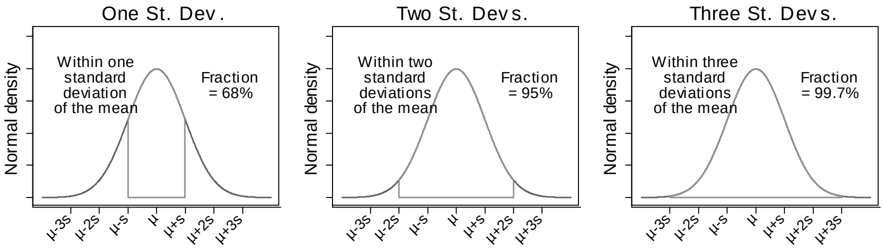
- **Sample variance**:

$$s^2 = \frac{\left[(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2\right]}{(n-1)} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

- The divisor $(n-1)$ is called the **degrees of freedom**
  - only $(n-1)$ terms in the sum can vary since the x_i are linked by $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.
- **Example:**
  - Sample $\{8, 3, 7, 6\}$ has $n = 4$ and $\bar{x} = 6$
  - $s^2 = \frac{1}{3}\left[(8-6)^2 + (3-6)^2 + (7-6)^2 + (6-6)^2\right] = 14/3 = 4.66.$
- **Sample standard deviation**: $s = \sqrt{s^2}$
  - Take square root to get back to same units as x.
  - e.g. Sample $\{8, 3, 7, 6\}$ has $s = \sqrt{s^2} = \sqrt{4.66} = 2.16.$

Interpretation of Standard Deviation

- Standard deviation is difficult to understand physically.
- As a guide use the fact that if data are normally distributed then 68%, 95%, 99.7% within 1, 2 and 3 standard deviations of the mean.
- And for any distribution at least 75% are within 2 standard deviations of the mean.



Data Dispersion or Spread: Other Measures

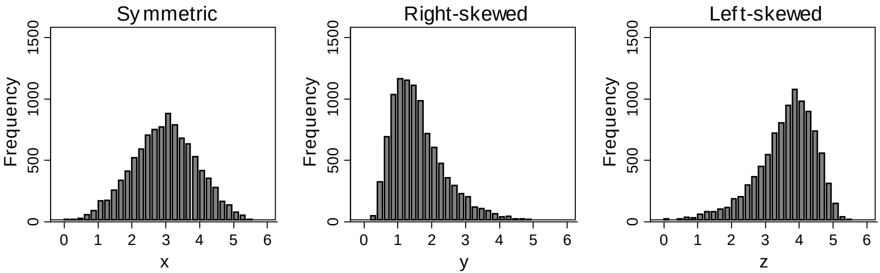
- We most often use the standard deviation.
- **Coefficient of variation:** $\text{CV} = s/\bar{x}$.
  - dispersion relative to the mean
- **Interquartile range**
  - difference between upper and lower quartiles.
- **Mean absolute deviation:** $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$
  - average absolute deviation about the mean.

Quartiles, Deciles and Percentiles

- These provide summaries of ordered data (in addition to the median).
- Quartiles split ordered data into fourths
    - **Lower quartile**: one-quarter below and three-quarters above
    - **Upper quartile**: three-quarters below and one-quarter above.
- Deciles split ordered data into tenths
    - **Ninth decile**: nine-tenths below and one-tenth above.
- Percentiles split ordered data into hundredths
    - **99th percentile**: 99% below and 1% above.

Skewness

- **Symmetry**
  - the density is the same when reflected about the mean
    - normal and t distributions are examples.
- **Skewness**: not symmetric.



- **Skewness statistic:** Approximately

$$\text{Skew} \simeq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3.$$

  - the average of the z-score $\left(\frac{x_i-\bar{x}}{s}\right)$ raised to third power
    - ★ where z-score is standardized to have mean 0 and variance 1
  - zero if no skewness
  - positive if right-skewed (e.g. prices, income) and negative if left-skewed.
- With skewed data mean $\neq$ median.
- For very skewed data may wish to use the median in addition to, or in place of, the mean.

Kurtosis

- **Kurtosis statistic:** Approximately

$$\text{Kurt} \simeq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 .$$

  - the average of the z-score $\left(\frac{x_i-\bar{x}}{s}\right)$ raised to fourth power.

- **Excess kurtosis** measures kurtosis relative to the normal distribution which has Kurt$= 3$

$$\text{ExcessKurt} = \text{Kurt} - 3.$$

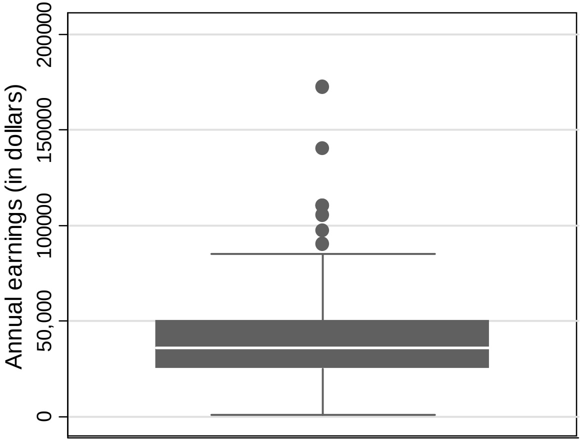
  - View **positive excess kurtosis** as fatter tails than normal.

- Since measure involves $(x_i - \bar{x})^4$ outliers get a lot of weight
  - financial returns data often have fat tails.

Box Plot

- A **box and whisker plot** or, more simply, a **box plot**
  - provides in a simple graphic some of the key summary statistics.
- Median is the middle line.
- Upper and lower quartiles are the lines surrounding the median.
- Outer bars vary with the statistical package
  - sometimes the minimum and maximum
  - sometimes the following is used to indicate outliers
    - upper bar is upper quartile plus 1.5 times the inter-quartile range
    - lower bar is lower quartile minus 1.5 times the inter-quartile range
    - dots are observations outside these bars.

Box Plot Example: Earnings



2.2 Charts for Numerical Data

- Standard charts for cross-section data are histogram and smoothed histogram.
- Example: Annual earnings of a sample of 171 female full-time workers aged 30 years in 2010
  - full-time is 35 or more hours per week and 48 or more weeks per year.
- The first nine observations are
  - 25000, 40000, 25000, 38000, 28800, 31000, 25000, 20000, 83000.
- Earnings range from $1,050 to $172,000.
- Earnings are generally reported to the nearest hundred or thousand or ten thousand dollars.

Frequency Distribution (tabulation in ranges)

- Summary of data grouped into intervals of width $15,000
  - e.g. 53 observations or 31% have earnings between $15,000 and $29,999.

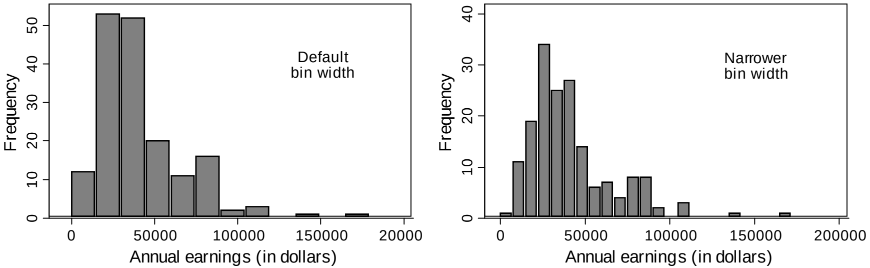
Range (or bin)	Frequency	Relative frequency (%)
0-14,999	12	7.0
15,000-29,999	53	31.0
30,000-44,999	52	30.4
45,000-59,999	20	11.7
60,000-74,999	11	6.4
75,000-89,999	16	9.4
90,000-104,999	2	1.2
105,000-119,999	3	1.8
120,000-134,999	0	0.0
135,000-149,999	1	0.6
150,000-164,999	0	0.0
165,000-180,000	1	0.6

Histogram

- The preceding table summarizes the data grouped into intervals of width $15,000
  - each interval is called a **bin**; here there are 13 bins $\simeq \sqrt{171}$.
  - each bin is of equal **bin width** of $15,000.
  - **frequency** is the number of observations that fall into a given bin
  - **relative frequency** is the proportion (or percentage) that fall into a given bin
- A **histogram** is a graph of the frequency distribution
  - horizontal axis: values or range of values
  - vertical axis: frequency or relative frequency or density (the relative frequency divided by the bin width)

Frequency Histogram for Two Bin Widths

- Smaller bin width gives more detail
  - here compare $15,000 to $7.500 bin width.

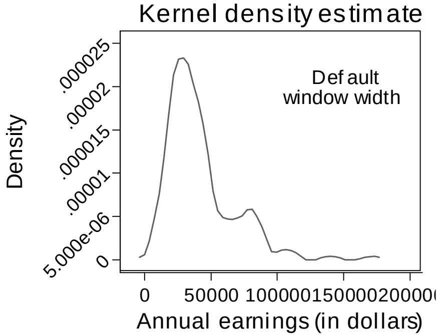


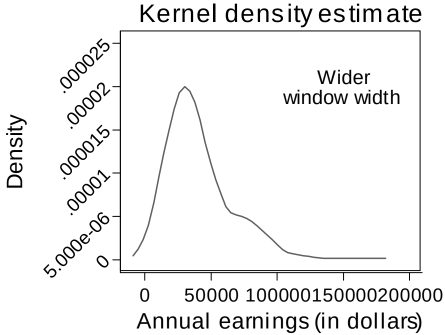
Smoothed Histogram (kernel density estimate)

- **Continuous data** such as earnings data have an underlying continuous density
  - e.g. the normal distribution (a bell-shaped density)
    - probabilities are determined by areas under the curve
    - total area under a density is one; see Appendix 5.A.
- A **kernel density estimate** is a commonly-used estimate of a density.
- It is a **smoothed histogram** that smooths in two ways
  - uses rolling bins (or **windows**) that overlap rather than being distinct
  - count the fraction of the sample within each bin with more **weight** given to observations at the window center and less to observations at the window ends.
- Can compare kernel density estimate to a proposed continuous density for the data such as normal.

Kernel Density Estimate for Two Window Widths

- Larger window width or bin width leads to smoother estimate.

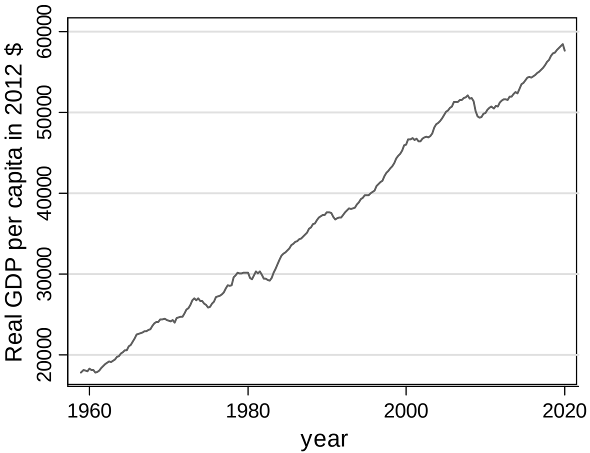




Line Chart for Ordered Data

- A standard chart for time series data is a line chart.
- A **line chart** plots the successive values of the data against the successive index values.
- Useful for numerical data where interest lies in how the data change from one observation to the next.
- Leading application is to time series data
  - these have a natural ordering of the observations, namely time.
- Next slide shows line chart for real gross domestic product (GDP) per capita in constant 2012 dollars from of data from 1959 to 2019.
  - indicates enormous improvement in living standards
    - ★ per capita real GDP tripled over the sixty years
  - also shows dips due to recessions.

Line Chart Example: Real per capita U.S. GDP



2.3 Charts for Numerical Data by Category

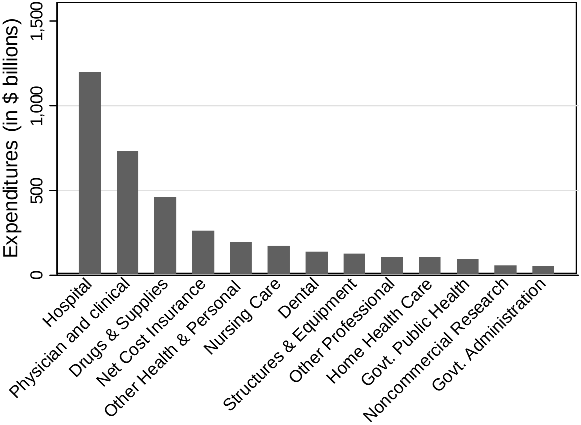
- U.S. health expenditures in 2018 of $3,653 billion (18% of GDP)
  - broken into its main subcomponents.

Category	Amount ($ billions)
Hospital Care	1192
Physician and Clinical Services	726
Dental	136
Other Professional	104
Other Health and Personal	192
Home Health Care	102
Nursing Care	169
Drugs and Supplies (Retail Sales)	456
Government Administration	48
Net Cost of Health Insurance	259
Government Public Health	94
Noncommercial Research	53
Structures and Equipment	122

Bar Charts

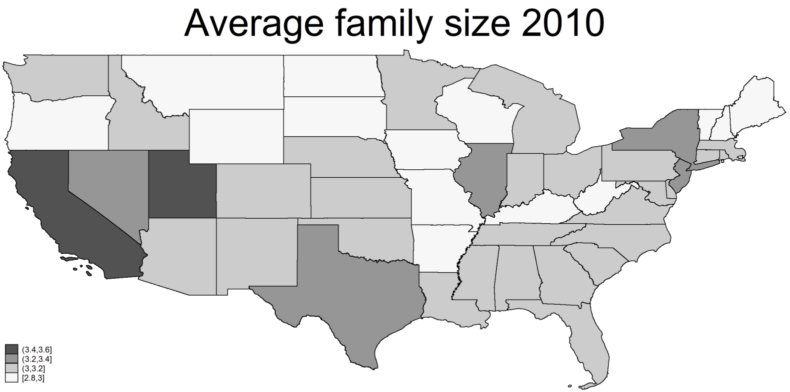
- Bar charts are a standard chart for numerical categorical data.
- A **bar chart**
  - provides a bar for each category
  - the length of the bar is determined by the category's value.
- A **column chart** or **vertical bar chart**
  - values on the vertical axis
  - category on the horizontal axis.
- A **horizontal bar chart**
  - values on the horizontal axis
  - category on the vertical axis.

Column Chart Example



Spatial Map

- Plot data by geographic location against a geographic map.
- Example is average family size in each U.S. state in 2010
  - darker shades correspond to larger families.



2.4 Summary and Charts for Categorical Data

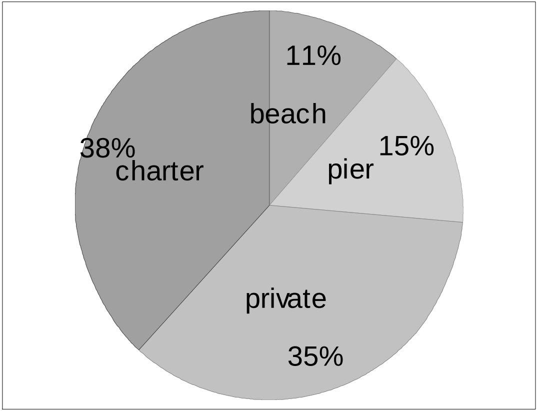
- Example: Fishing site chosen by a sample of 1,182 fishers
  - there are four possible sites (categories).
- Summarize using a Tabulation of frequencies.

Category	Frequency	Relative frequency (%)
Beach	134	11.34
Pier	178	15.06
Private Boat	418	35.36
Charter Boat	452	38.24

Pie Chart

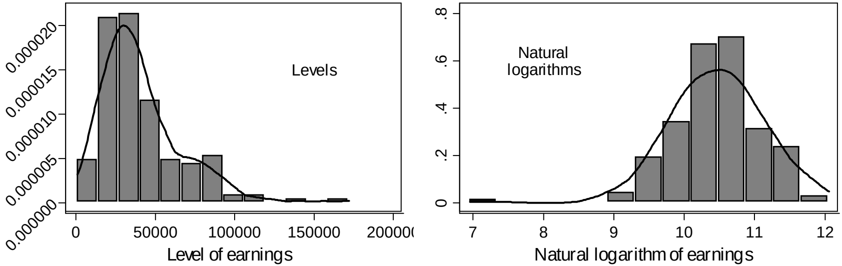
- A **pie chart** splits a circle into slices
  - the area of each slice corresponds to the relative frequency of observations in each category.
- A pie chart with many categories can be made easier by
  - giving the slices in order of decreasing size
  - giving the associated headings, in the same ordering, in a separate legend.
  - using color rather than black-and-white.

Pie Chart Example



2.5 Data Transformations: Natural Logarithm

- Many economic series are right-skewed: prices, income, wealth, ...
  - natural logarithm converts right-skewed data to a more symmetric distribution.



- Advantages of using natural logarithm are given in Chapter 9.

Standardized Scores

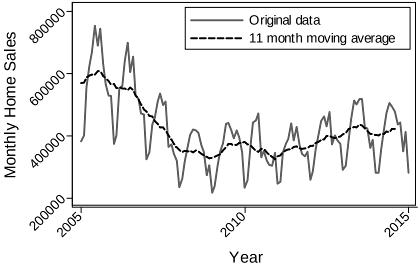
- Standardized scores (or z-scores)
  - Consider sample with sample mean $\bar{x}$ and standard deviation s
  - subtract the mean and divide by the sample standard deviation
  - so

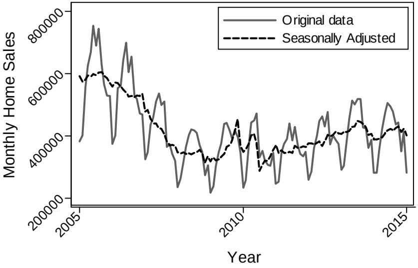
$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, ..., n$$

- Then $z_1, ..., z_n$ has mean $\bar{z} = 0$ and sample standard deviation one.
- Useful for comparing series that are scaled differently
  - e.g. test scores on two different tests.
- If e.g. $z_i = -3$ then x_i was 3 standard deviations below the mean.

2.6 Common Data Transformations for Time Series Data

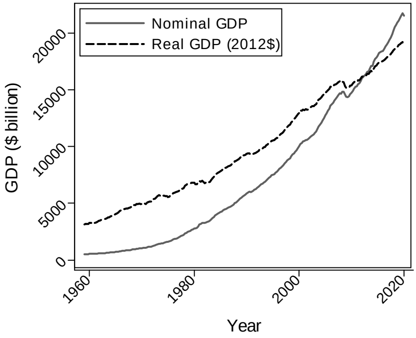
- Moving averages: smooth by averaging over several successive periods.
- Seasonal adjustment: smooth by adjusting for seasonal variation.

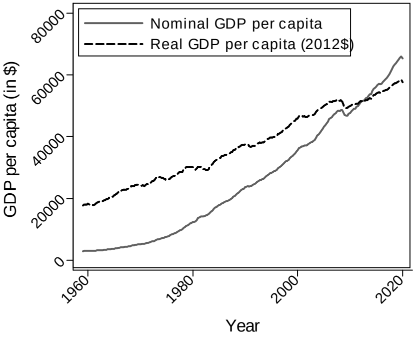




More Data Transformations for Time Series Data

- Real and nominal data: adjust for price inflation.
- Per capita data: adjust for population size.





Growth rates and percentage changes

- The one-period **percentage change** in x_t is $100 \times \frac{x_t - x_{t-1}}{x_{t-1}}$.
  - this is often converted to an annualized rate
    - e.g. for quarterly data the quarterly change is multiplied by four.
- Distinguish between **percentage point change** and percentage change
  - Suppose the growth rate increases from 3 percent to 5 percent
  - correct: the growth rate increased by two percentage points
  - incorrect: there is a 2 percent increase in the growth rate
    - which means an increase from 3.0 percent to $3.0 \times 1.02 = 3.06$ percent.
- Very small changes are described in **basis points**
  - a basis point is one-hundredth of a percentage point.
- An approximation (explained in chapter 9.1) is
  - proportionate change in x = level change in natural log of x
  - so percentage change in $x_t \equiv 100 \times \frac{x_t - x_{t-1}}{x_{t-1}} \simeq 100 \times (\ln x_t - \ln x_{t-1})$.

Key Stata Commands

```
clear
use AED_EARNINGS.DTA
describe
summarize
list earnings in 1/5
summarize earnings
summarize earnings, detail
histogram earnings, freq
kdensity earnings
histogram earnings, kdensity
generate lnearns = ln(earnings)
kdensity lnearns, normal
```

Some in-class Exercises

1. Obtain $\sum_{i=1}^{3}(2+3i^2)$.
2. Obtain the mean, variance and standard deviation for a sample with values 5, 2, 2.
3. For a sample of size 500 from the normal distribution, approximately how many observations do you expect to be within two standard deviations of the mean?
4. For a sample with mean 3 and variance 4 find the z-score for an observation with value 6.
5. If x increases from 4 to 5 what is the percentage change in x?