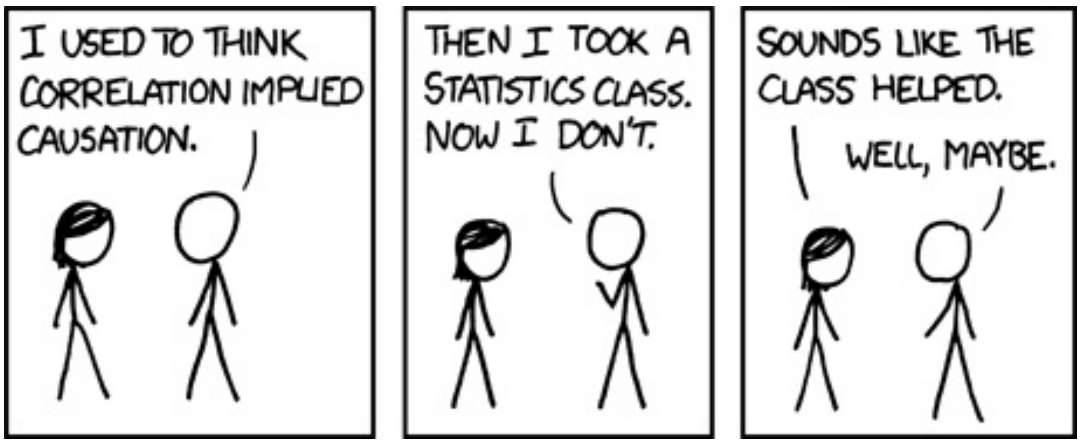
Introduction

ECN 102: Analysis of Economic Data

Winter, 2011



Contact Information

- Instructor: John Parman
  - Email: jmparman@ucdavis.edu
  - Office: 1125 SSH (NW entrance to building)
  - Office hours: Monday and Thursday, 2pm - 4pm
- TAs:
  - Kuk Mo Jung (kmjung@ucdavis.edu)
  - Danielle Sandler (dhsandler@ucdavis.edu)
  - Yi Chen (yiychen@ucdavis.edu)

Course Website

- We will have a course website on Smartsite:

  smartsite.ucdavis.edu

- The syllabus, problem sets, past exams, solutions, data files and grades will all be posted there
- Lecture slides will be posted, typically about 30 minutes before lecture
- If you are open campus or auditing the course, let me know and I will give you access to the Smartsite page

Textbook

- The required text is *Analysis of Economic Data* by Colin Cameron.
- It is available as a course reader from Davis Textbooks (3rd and A).
- You can use older versions of the reader.
- There will be a copy on reserve in the library.

Waitlist, PTA numbers, etc.

- The course is currently full.
- The only way to get into the course is through the waitlist, no PTA numbers will be given.
- For open campus students, you can't be enrolled until after the drop/add period is over. In the meantime, send me an email and I will give you access to the Smartsite page so that you can keep up with the course.

Grading

Grades will be based on problem sets, two midterms and a final exam, weighted as follows:

Problem Sets:	10%
Midterm 1:	25%
Midterm 2:	25%
Final:	40%

Grades for the course will be curved such that the average GPA for the course is a 2.4. Although the curve will be based on the distribution of overall course grades at the end of the quarter I will give you a rough idea after each exam of what letter grades correspond to different ranges of the uncurved numerical scores.

Schedule

Week of	Tuesday	Thursday
January 3	lecture	lecture
January 10	lecture	lecture
January 17	lecture	lecture
January 24	lecture	**Midterm 1**
January 31	lecture	lecture
February 7	lecture	lecture
February 14	lecture	lecture
February 21	lecture	**Midterm 2**
February 28	lecture	lecture
March 7	lecture	lecture

Final Exam: Thursday March 17, 10:30am-12:30pm

Exams

- All exams will be cumulative but will place greater emphasis on new topics (I will go over what that means closer to the exams).
- For each exam, you will need to bring a scantron sheet (UCD 2000), something to write with and a non-graphing calculator.
- You have one week after any graded material is returned to raise any grading issues. You must submit regrade requests in writing and include an explanation of why a regrade is warranted.

Problem Sets

- Problem sets will be posted online and announced in class. Four of the problem sets will be collected and graded. It will state on the problem set whether or not it will be graded.
- Grading will be on a check plus, check, check minus scale.
- You may work in groups on problem sets but each person must write up and submit his or her own problem set. This includes creating your own tables, graphs, etc.
- Problem sets will typically involve a fair amount of work in Excel (learning how to use the 'set print area' function will be very useful).

Excel, data, etc.

- You will often have to use Excel and data provided on the course website for problem sets. Excel 2007 will be used in class and in sections to demonstrate how to work with data.
- You may use other versions of Excel or other programs (OpenOffice, Stata, etc.) to do the homework. Datasets will be provided in a generic format so that it can be used in whichever program you choose.
- Excel 2007 and Stata are available on the lab computers in Hutchison.
- Helpful handhouts on using Excel can be found on Professor Cameron's website:

  http://cameron.econ.ucdavis.edu/excel/excel.html

Uses of Economic Data

- To describe the economic "landscape"
  - Examples:
    - What is the annual growth rate of GDP? Has unemployment risen over the past year?
    - Do people with higher levels of education tend to have greater earnings? Do democracies have greater growth rates than dictatorships?
  - Descriptive statistics motivate economic theory
- To test or attempt to distinguish between economic theories
- To help guide policy and expectations about the future

Types of Data

There are a variety of different types of data that you will encounter in economics. The ways in which we categorize types of data include the following:

- Value: numerical data, categorical data
- Unit of observation: cross-section data, time series data, panel data
- Number of variables: univariate data, bivariate data, multivariate data

Types of Data: Numerical Data

Numerical data are data that are naturally recorded and interpreted as numbers. They can be continuous or discrete. Examples of numerical data include:

- Annual income (continuous)
- Hours worked (discrete)
- Annual GDP (continuous)
- Number of times a person has moved (discrete)

Types of Data: Categorical Data

Categorical data are data that are recorded as belonging to one or more groups. They can be recorded as numbers but these numbers have no inherent meaning. Examples of categorical data include:

- Gender
- Birthplace
- Religion

Types of Data: Cross-section Data

Cross-section data are data on different individuals collected at a common point in time.

- Notation: x_i, $i = 1, ..., n$
  - i specifies a particular individual for an observation
  - n is the total number of individuals observed (typically called the sample size)
  - x is the value of whatever variable we are observing
- Examples: a single year of census data, GDP by country for a particular year, unemployment rates by state for a particular year

Types of Data: Time-Series Data

Time-series data are data on a particular phenomenon collected at different points in time.

- Notation: x_t, $t = 1, ..., T$
  - t specifies the time period of an observation
  - T is the total number of time periods
  - x is the value of whatever variable we are observing
- Examples: GDP over time, daily averages of the S & P 500, monthly unemployment rates

Types of Data: Panel Data

Panel data are data on different individuals with each individual observed at multiple points in time.

- Notation: $x_{i,t}$, $i = 1, ..., n$; $t = 1, ..., T$
- Panel data is a mixture of cross-section and time series data
- Examples: Earnings of Davis graduates over time, life expectancy by country over time

Types of Data: Univariate Data

Univariate data is a single data series containing observations of only one variables.

- Notation: x_i for cross-section data, x_t for time series data
- Examples: Earnings of high school graduates in 2008, inflation rate from 1950 to 2008

Types of Data: Bivariate Data

Bivariate data is composed of two potentially related data series.

- Notation: (x_i, y_i) (cross-section data), (x_t, y_t) (time series data)
- We're often interested in the relationship between x and y
- Examples: education and earnings for high school graduates, inflation and unemployment rates over time

Types of Data: Multivariate Data

Multivariate data is composed of three or more potentially related data series.

- Notation: $(x_{1,i}, x_{2,i}, ..., x_{K,i}, y_i)$ (cross-section data), $(x_{1,t}, x_{2,t}, ..., x_{K,t}, y_t)$ (time series data)
- We're often interested in how $x_1, ..., x_K$ are related to y
- Examples: inputs, outputs and profits for a firm over time; education, gender and income for a cross-section of individuals

What do we do with economic data?

The basic steps of data analysis:

1. Data summary
2. Statistical inference
3. Interpretation

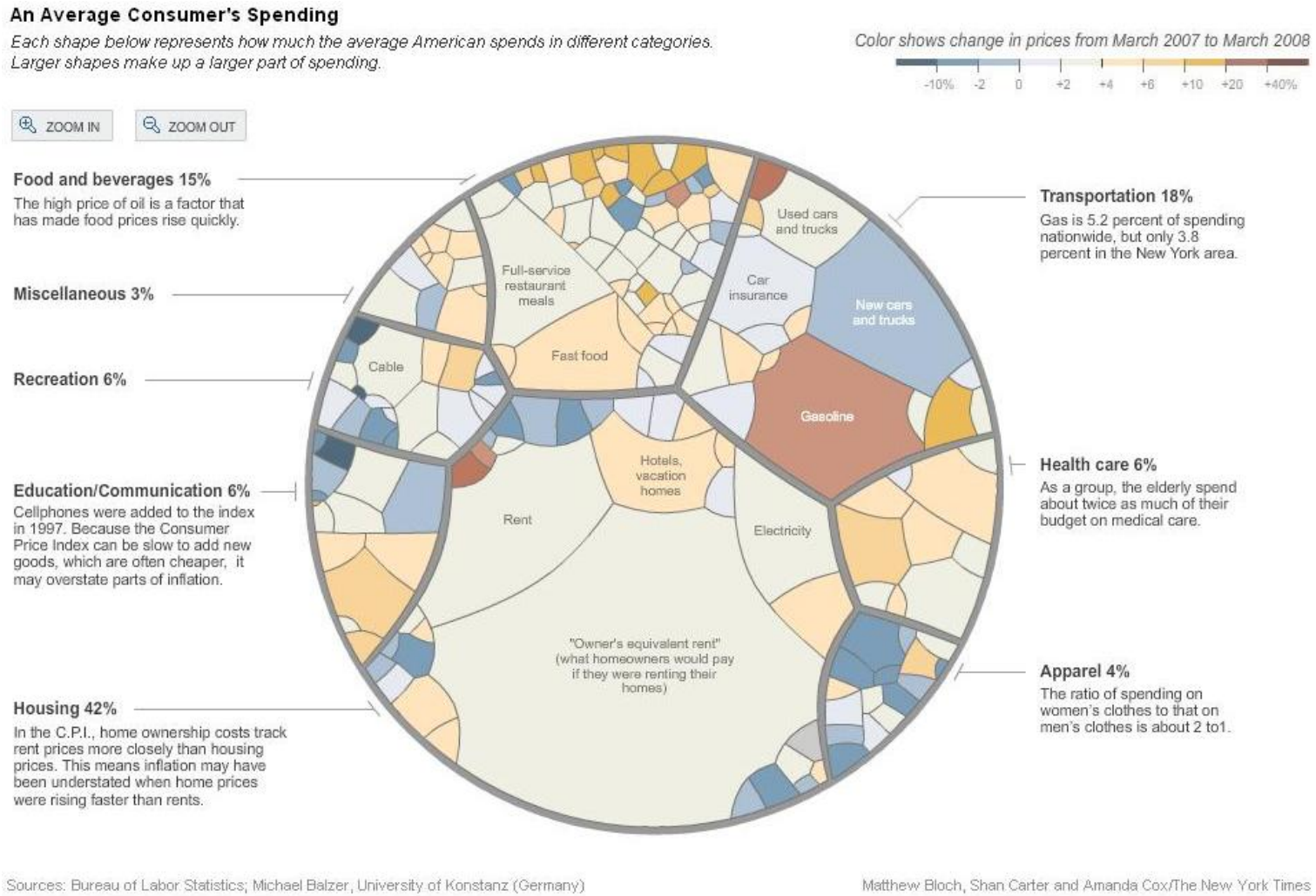
Steps of Data Analysis: Data Summary

- To summarize data, we typically use a combination of visual representations of the data and statistics
- Visual representations include a variety of graphs and charts (scatterplots, histograms, maps, etc.)
- Statistics can measure characteristics of a single variable (mean, median, variance, etc.) or relationships between multiple variables (covariance, correlation, linear regression, etc.)
- The choice of summary statistics and graphs depends on both the type of data available and what the researcher is interested in

Steps of Data Analysis: Statistical Inference

- The basic idea of statistical inference is to draw conclusions about a relationship we cannot observe
- We typically cannot reach definitive conclusions because we only get to observe a *sample* rather than the *population*
- Statistical inference requires using what we know about the sample and about probability to reach a conclusion about the probable characteristics of variables and relationships between them at the population level

Graphical Representations of Univariate Data



Summary Statistics for Univariate Data

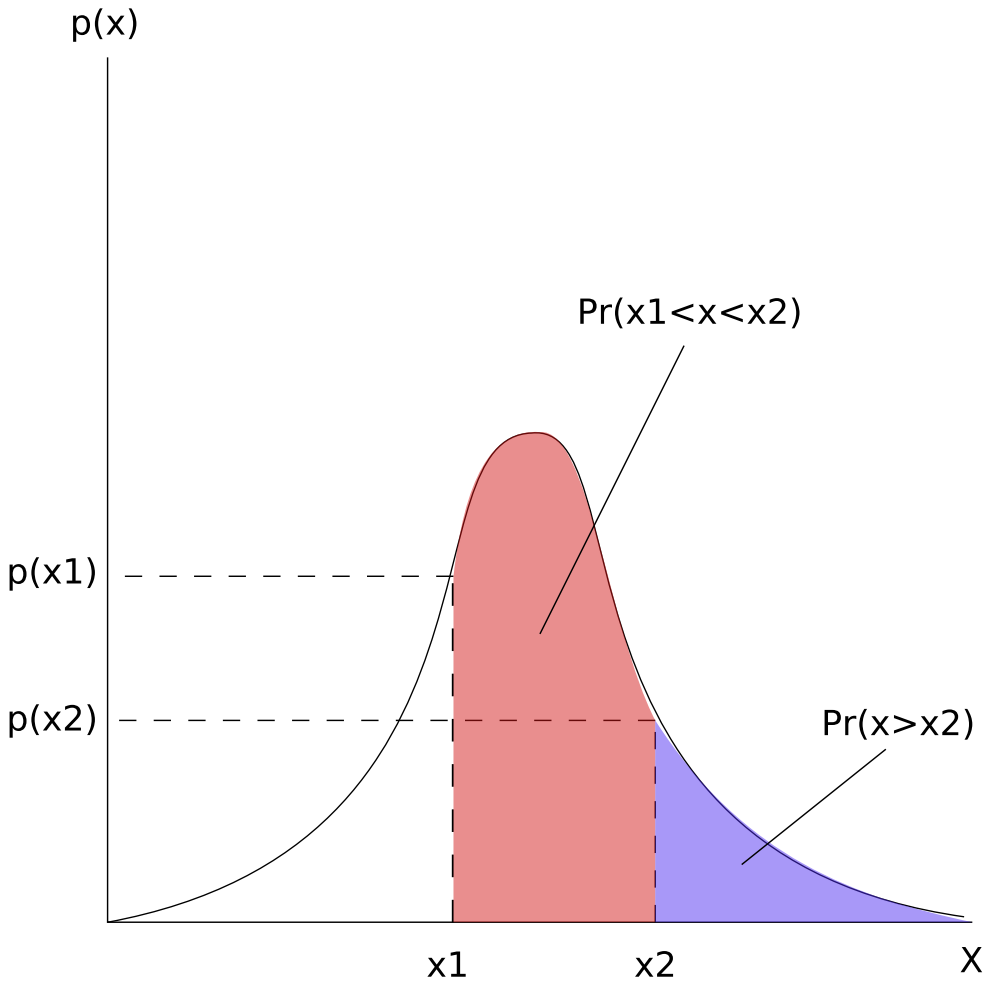
- Graphs are nice for giving people a quick glimpse of data
- However, there is a lot of ambiguity about interpreting graphs and comparing one to another
- Where is the mean? What is a wide distribution and what is a narrow one? Are tails big or small? Etc.
- Summary statistics give us a standardized way of summarizing univariate data
- People know what the numbers mean and they can be compared across different samples

A Little Stats Review

What do we mean by a distribution?

- Probability distributions
- Frequency distributions
- Sample vs. population

A Little Stats Review



A Little Stats Review

The summation operator:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

If a and b are constants:

$$\sum_{i=1}^{n} a = n \cdot a$$

$$\sum_{i=1}^{n} b \cdot x_i = b \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

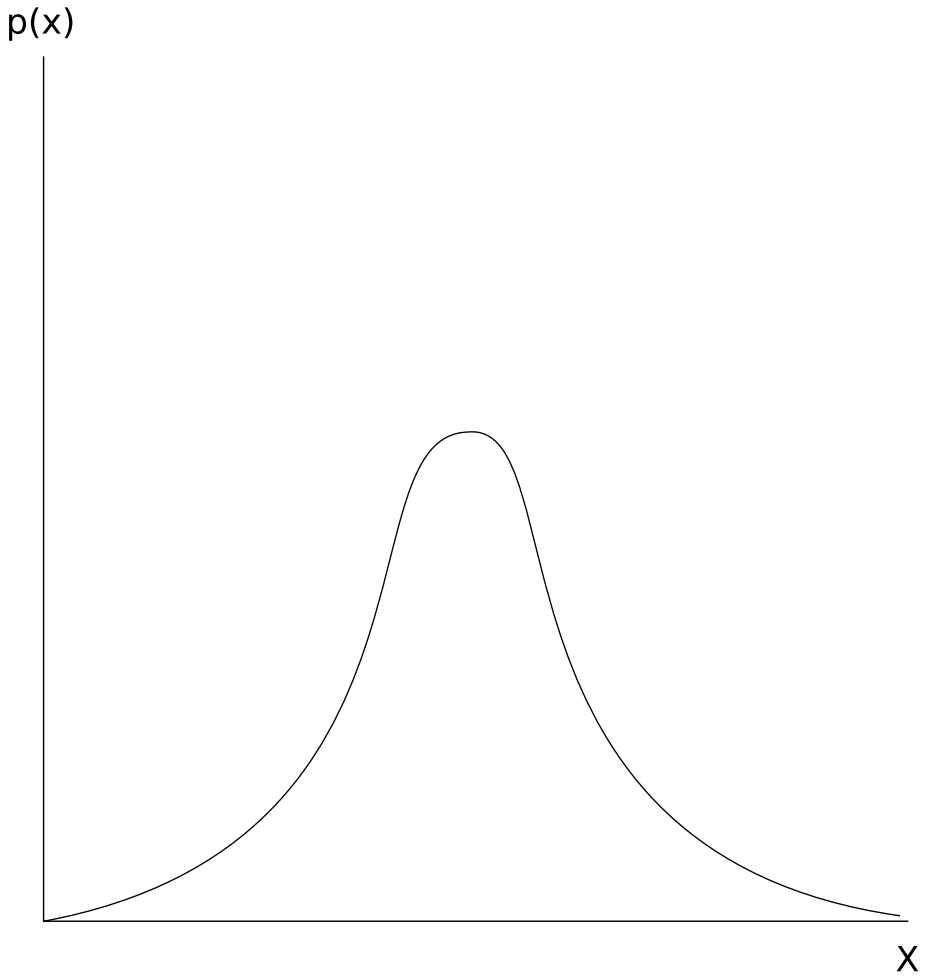
$$\sum_{i=1}^{n} (x_i \cdot y_i) \neq (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)$$

Types of Summary Statistics

We're often interested in describing the following characteristics of the distribution of a data series:

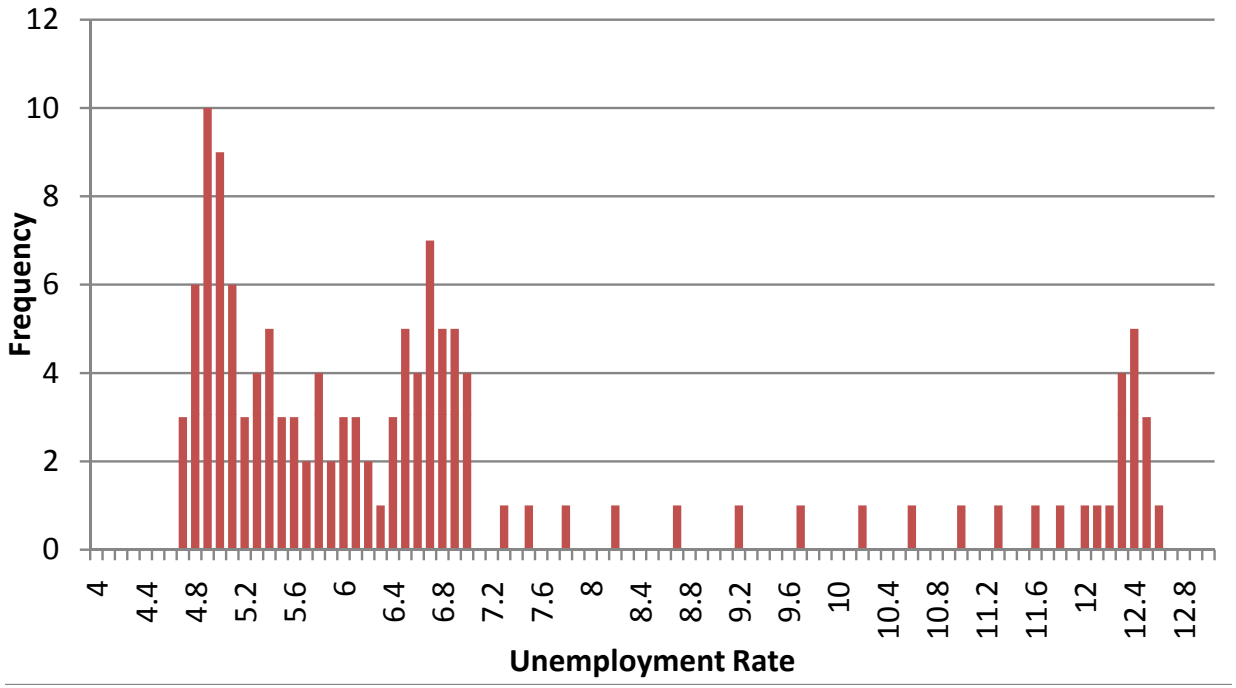
- **Central tendency** - where is the middle of the distribution?
- **Dispersion** - how spread out is the data?
- **Skewness (asymmetry)** - how symmetric (or assymetric) is the distribution?
- **Peakedness** - how fat are the tails, how tall is the peak?

Types of Summary Statistics



Types of Summary Statistics

To go over these different types of summary statistics, we'll use the following example:



This is the distribution of monthly unemployment rates for California for the past 10 years. The data are in ca-urate-2000-2010.csv.

Measures of Central Tendency

- Tells us where center of distribution is
- Answers the question, "What is a typical value in this sample?"
- Several different measures:
  - Sample average (sample mean)
  - Sample median
  - Sample midrange
  - Sample mode

The Sample Average

- Most common way to measure central tendency
- Definition:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Weights all observations equally
- Excel: use AVERAGE() formula

The Sample Median

- Value that divides the sample into two halves (50% of observations are above value and 50% are below)
- When n is an odd number, median is the middle value, when n is an even number, use the average of the two middle observations
- Less sensitive to outliers than the sample average
- Other quantiles can be used
- Excel: use MEDIAN() formula (PERCENTILE() for other quantiles)

The Mean Vs. The Median

- The mean household income in Medina, WA:

  $257,258

- The median household income in Medina, WA:

  $169,196

- Note that the mean is over 50% larger than the median
- Why is there such a big difference? Which of these numbers is more relevant?

The Sample Midrange

- The sample midrange is the average of the smallest and largest observations
- Not a very commonly used measure
- Extremely sensitive to outliers
- Excel: use the MIN() and MAX() functions:

$$= \frac{1}{2}(MIN() + MAX())$$

The Sample Mode

- The most frequently occurring value in sample
- Useful with discrete data and cases where particular values are meaningful (4 years of high school,40 hours of work each week, ...)
- Excel: use MODE() function

Measures of Dispersion

- Characterize the spread or width of the distribution
- Different measures:
  - Sample variance
  - Sample standard deviation
  - Sample coefficient of variation
  - Sample range and inter-quartile range
- Like measures of central tendency, the different measures have different benefits and drawbacks

Sample Variance

- Approximately equal to the average squared deviation from mean:

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

- As the sample variance increases, the spread of the data gets wider
- Excel: use VAR() function

Sample Standard Deviation

- Standard deviation is just the square root of the variance:

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

- Roughly the average deviation of the data from its mean
- Has the same units as the data
- Excel: use STDEV() function

Sample Coefficient of Variation

- Sample standard deviation relative to sample mean:

$$CV = \frac{s}{\bar{x}}$$

- Standardized measure: no units, can be compared across series
- Excel: use both the STDEV() function and the AVERAGE() function

$$= STDEV()/AVERAGE()$$

Sample Range

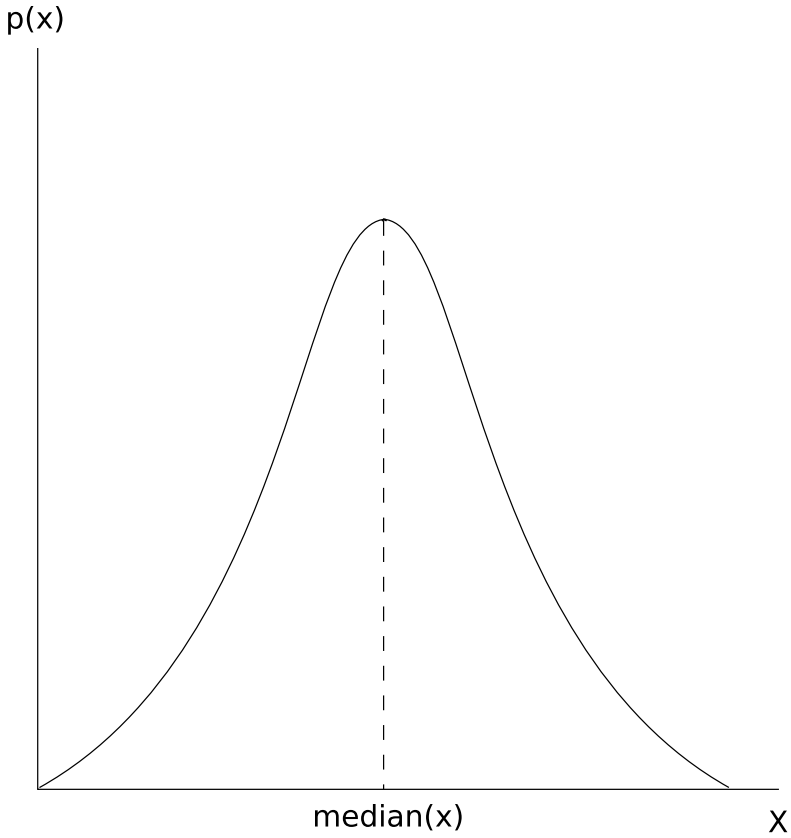
- Difference between the largest and smallest values in the sample
- Simplest measure of dispersion but also the least interesting
- Very sensitive to outliers
- Excel: MAX() minus MIN()

Sample Inter-Quartile Range

- Variation on sample range that is less sensitive to outliers
- Equal to difference between 75th and 25th percentile of the distribution
- Excel: PERCENTILE(,.75)-PERCENTILE(,.25)
- Can use other percentiles as well

Symmetric Distributions

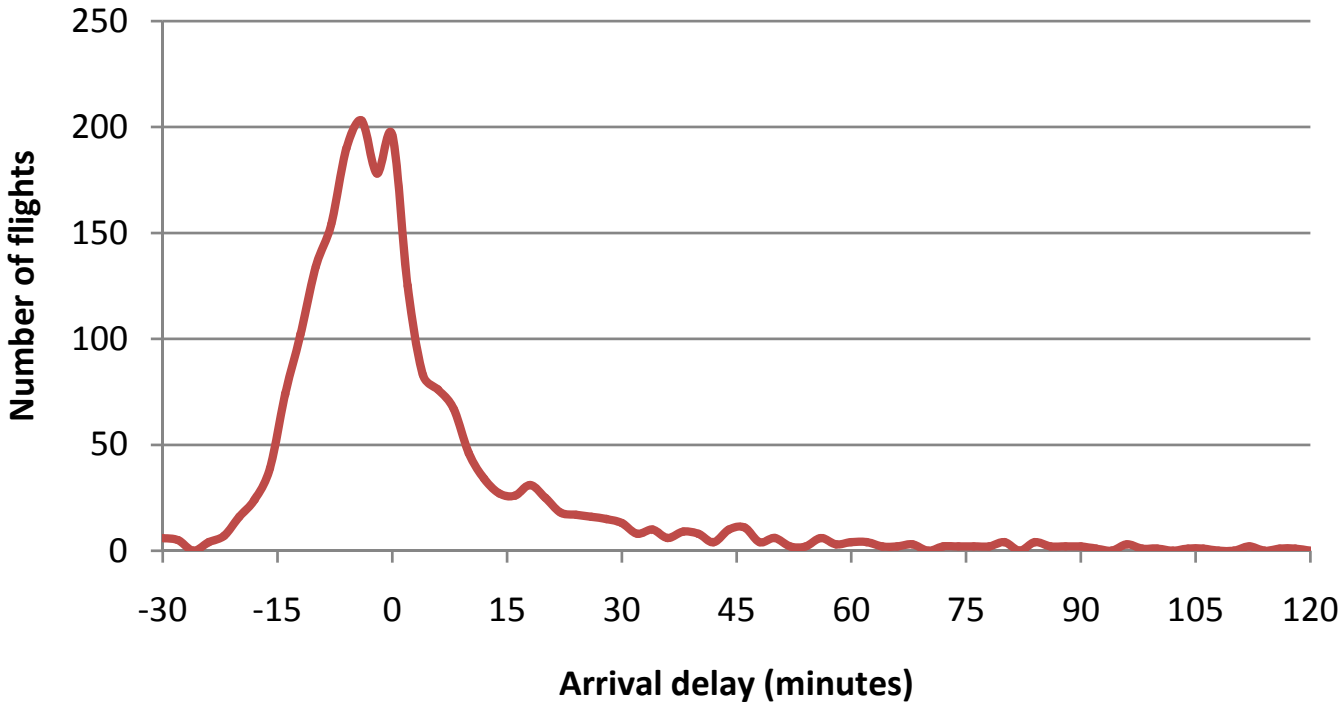
A distribution is symmetric if its shape is the same when reflected around the median



Measuring Symmetry (or Asymmetry)

- Typically use skewness to measure symmetry
- Right-skewed: distribution has a long right tail and data are concentrated to the left
- Left-skewed: distribution has a long left tail and data are concentrated to the right
- One way to test for right- or left-skewed is to compare median to mean:
  - Symmetric: $\bar{x} = median(x)$
  - Right-skewed: $\bar{x} > median(x)$
  - Left-skewed: $\bar{x} < median(x)$

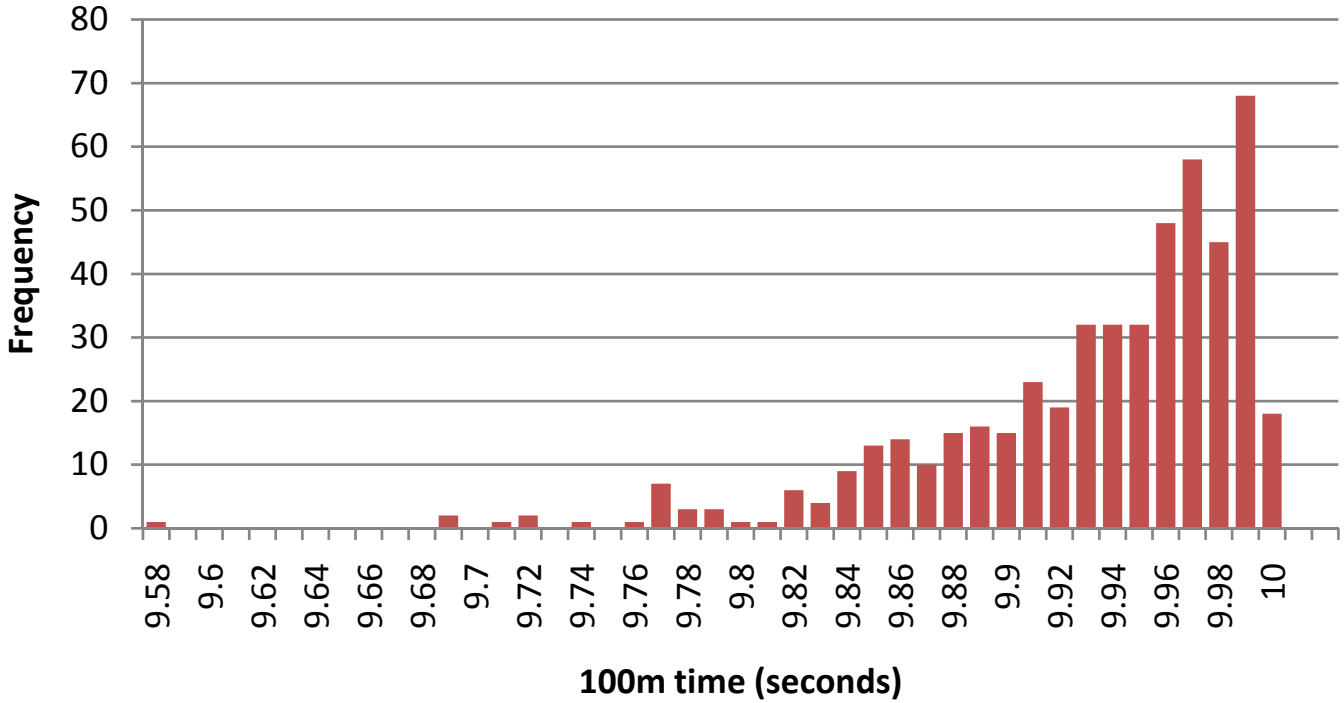
A Right-Skewed Distribution



Distribution of arrival delays for Southwest flights into SMF, January 2010

Mean = 3.4 min , Median = -2 min , Skewness = 5.0

A Left-Skewed Distribution



Distribution of the 500 fastest 100m times as of December 2010

Mean = 9.93 sec , Median = 9.95 sec, Skewness = -1.6

Quantifying Skewness

- The basic idea is to compare the mean with the median
- How we actually do it:

$$\frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- Interpretation of statistic: 0 if symmetric, greater than 0 if right-skewed, less than zero if left skewed
- Excel: use SKEW() function

Measuring "Peakedness"

- Peakedness is a question of how fat the tails of a distribution are
- Formally, we use kurtosis:

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

- Excel: use KURT() function

Interpreting Kurtosis

- Kurtosis has no units (because $x_i - \bar{x}$ is divided by s)
- If kurtosis is equal to 0, the distribution has the shape of the normal distribution
- If kurtosis is greater than 0, the distribution is peaked relative to the normal distribution and has fat tails
- If kurtosis is less than 0, the distribution is less peaked relative to the normal distribution and has skinny tails

Interpreting Kurtosis