

Lec1-1, Spring, 2005

# STAT 220E, Spring, 2005

# Displaying Distributions with Graphs

## Outline

- Individuals and Variables
- Categorical and Quantitative Variables
- Exploratory data analysis
- Categorical variable analysis
  - Pie charts.
  - Bar graphs
  - Pareto charts
- Quantitative variable analysis
  - Stemplots
  - Histograms

## Some Definitions

- **Individual:** each object described by a set of data
- **Variable:** any characteristic of an individual
  - **Categorical variable:** places an individual into one of several groups or categories.
  - **Quantitative variable:** takes numerical values on which we can do arithmetic.
- **Distribution of a variable:** tells what values it takes and how often it takes these values.

### Example

	Name	Age	Gender	Race	Salary	Job type
1	F Delores	39	Female	White	62,100	Management
2	P Juan	27	Male	White	47,350	Technical
3	W Lin	22	Female	Asian	18,250	Clerical
4	J LaVerne	48	Male	Black	77,600	Management

# Categorical variable analysis

Questions to ask about a categorical variable:

1. How many categories are there?
2. In each category, how many observations are there?

## Pie Charts

In **pie chart**, the area of each slice is proportional to percentage of that category among the **whole** population you are considering.

Attention: You need to be clear about what the whole population is!

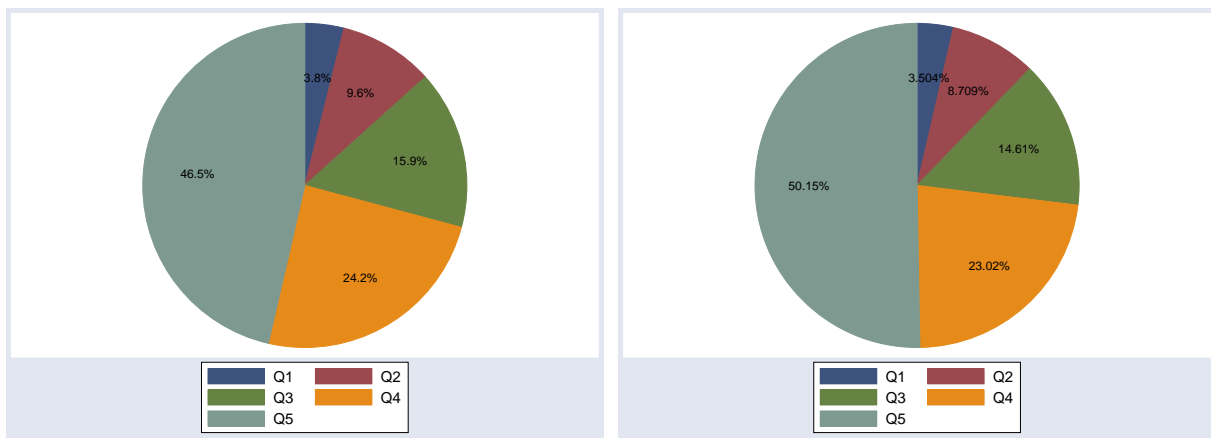
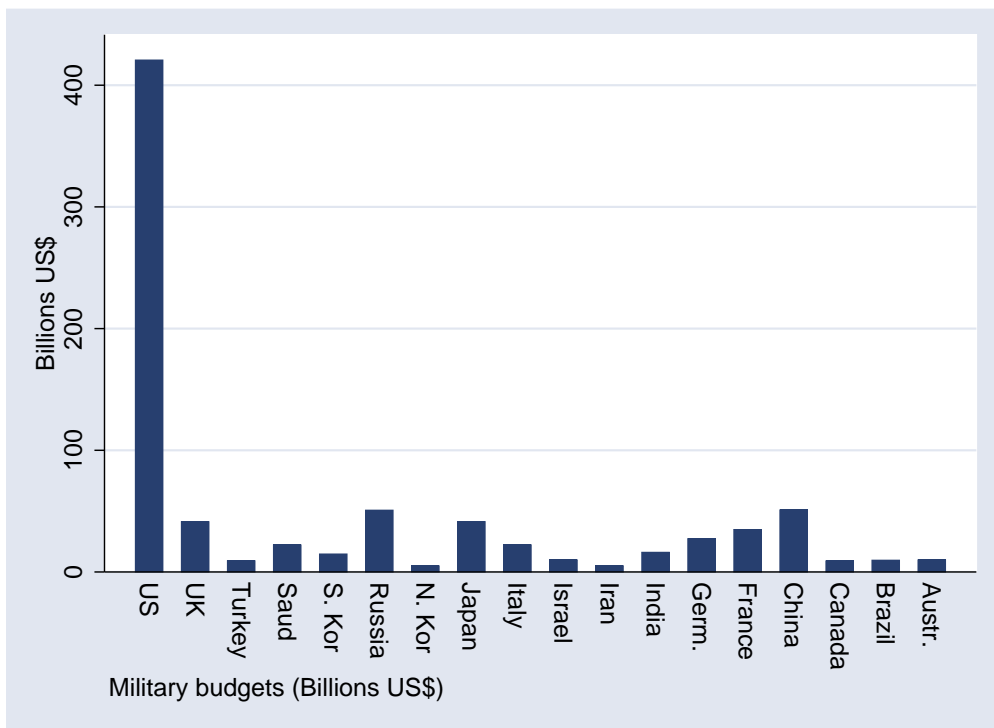


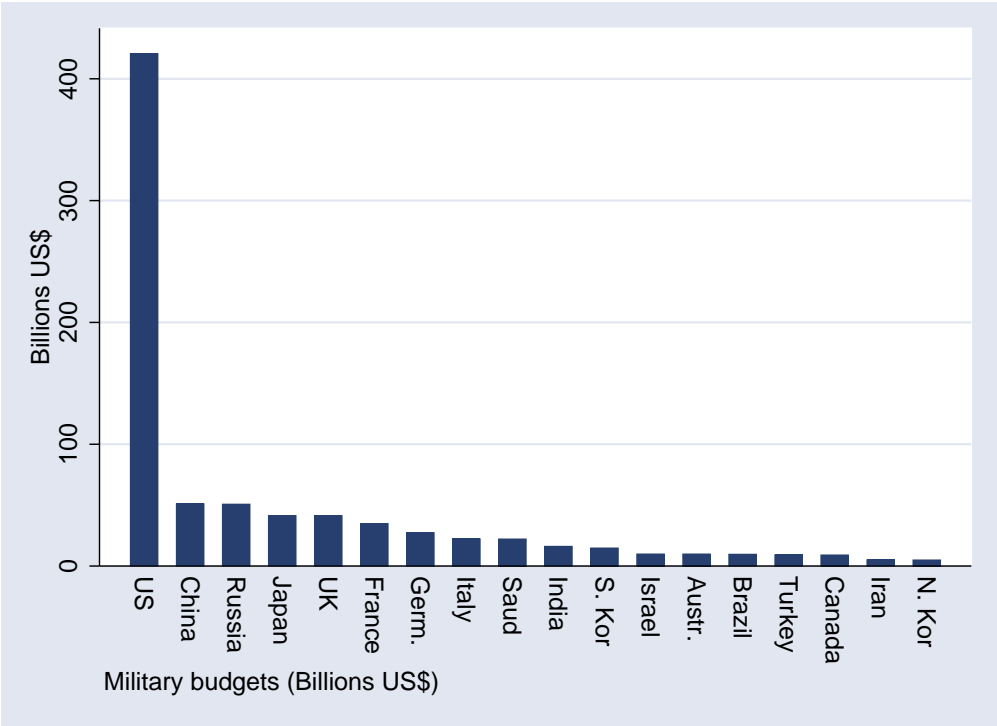
Figure 1: Share of total income of US household by quintiles. Left 1991, Right 2001

## Bar Graphs

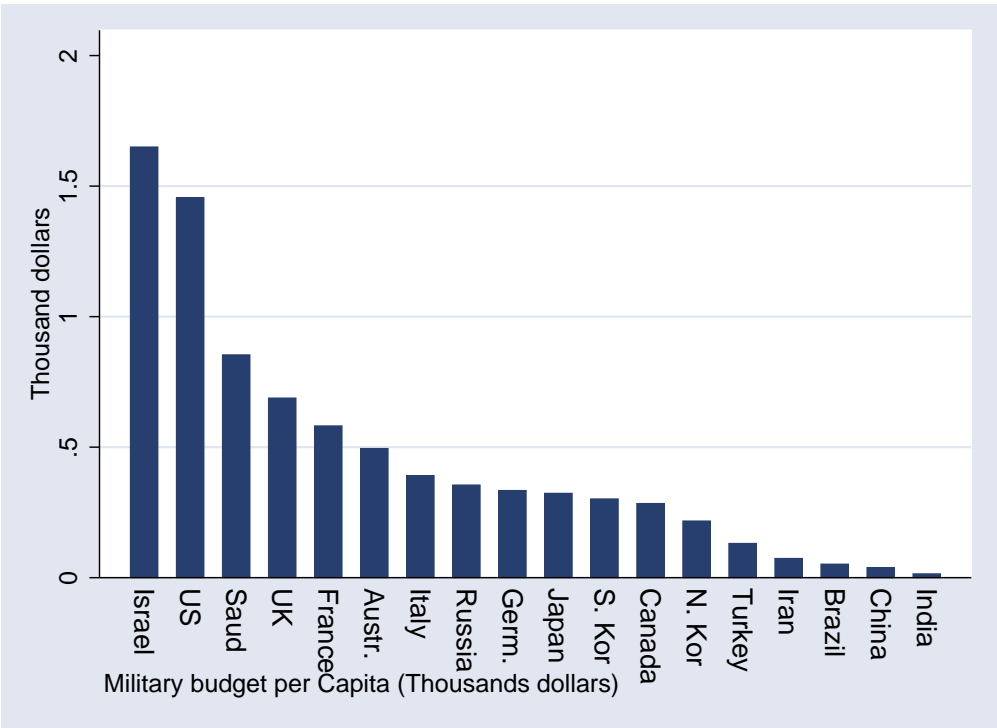
In a **bar graph**, the horizontal axis lists the categories, in any order. Bar height can be either counts or percentages of categories.



A **Pareto chart** is a special bar graph with categories ordered from most to least frequent.



Different relations between variables may be more interesting:



# Quantitative variable analysis

## Stemplots

### How to make a stemplot

- (1) Separate each observation into a stem and a leaf.

$$\text{e.g. } 265 \rightarrow \underbrace{26}_{\text{stem}} \underbrace{5}_{\text{leaf}}$$

- (2) Write the stems in a vertical column in increasing order.
- (3) Go through the data writing each leaf on the proper stem.
- (4) Arrange the leaves on each stem in order out from the stem.

## Example: making a stemplot

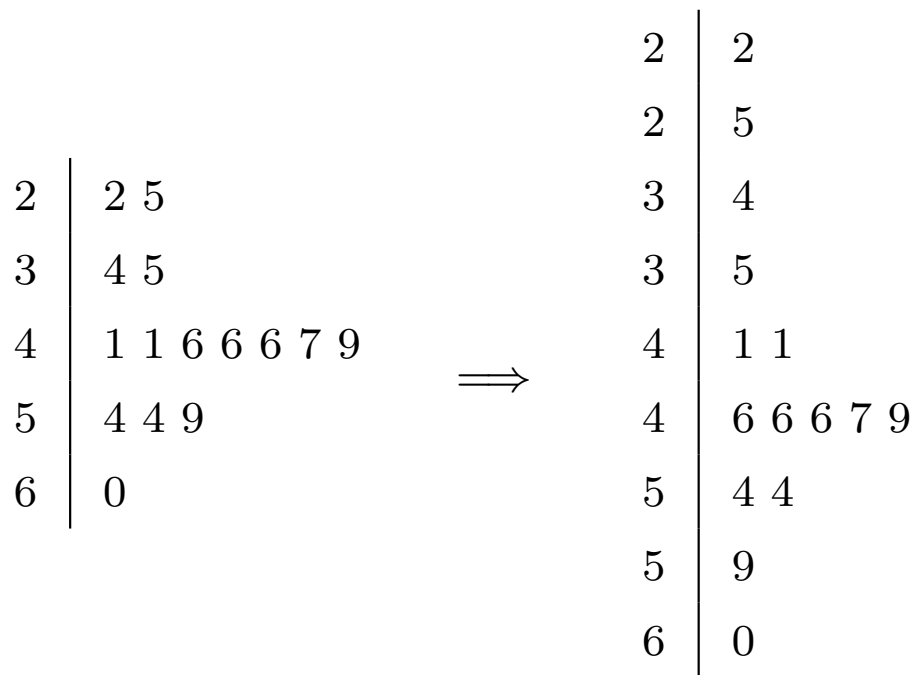
Here are the number of home runs that Babe Ruth hit in each of his 15 years with the New York Yankees, 1920 to 1934.

54	59	35	41	46
25	47	60	54	46
49	46	41	34	22

(From *The Baseball Encyclopedia*, 3rd ed., Macmillan, New York, 1976)

2		2	5 2	2	2 5
3		3	5 4	3	4 5
4		4	1 6 7 6 9 6 1	4	1 1 6 6 6 7 9
5		5	4 9 4	5	4 4 9
6		6	0	6	0
(2)		(3)		(4)	

## Example: splitting a stemplot



# Histograms

## How to make a histogram

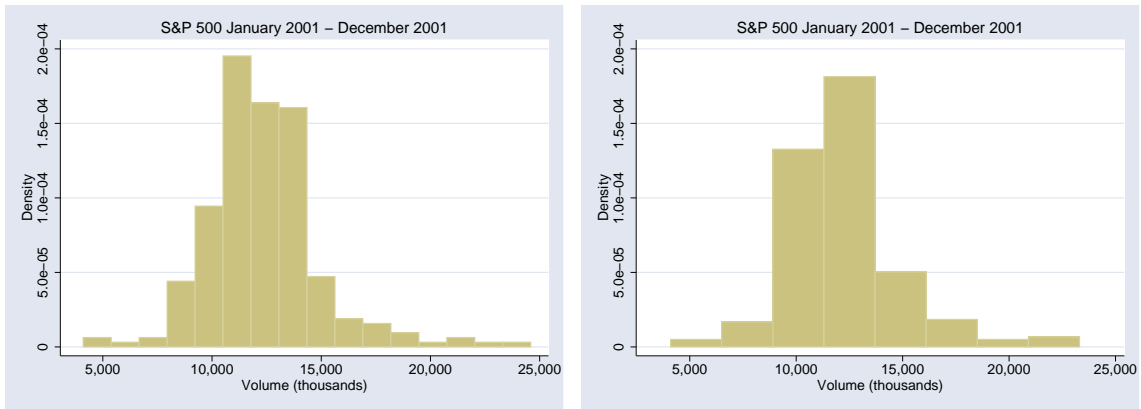
1. Group the observations into “bins” according to their value. Choose the bins carefully: too few hide detail, while too many decimate the pattern.
2. Count the individuals in each bin.
3. Draw the histogram
  - Leave no space between bars.
  - Label the axes with units of measurement.
  - The  $y$ -axis can be counts or percentages.

*The area of each bar is proportional to the percentage of data in that range. We care about the **area**, not the **height**, but when the bar has equal width, area is determined by the height.*

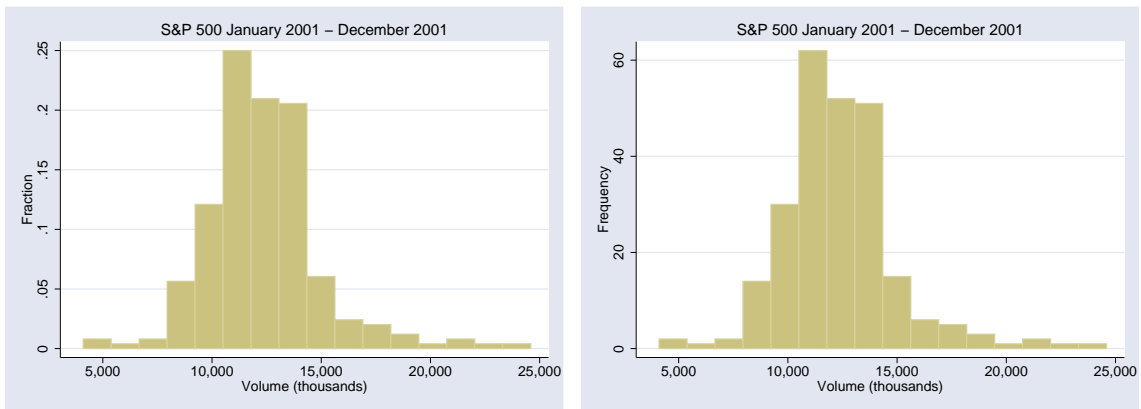
For simplicity, **use equally spaced bins.**

# Example

Volume of shares traded for S&P500 stocks.



Denisty histograms.



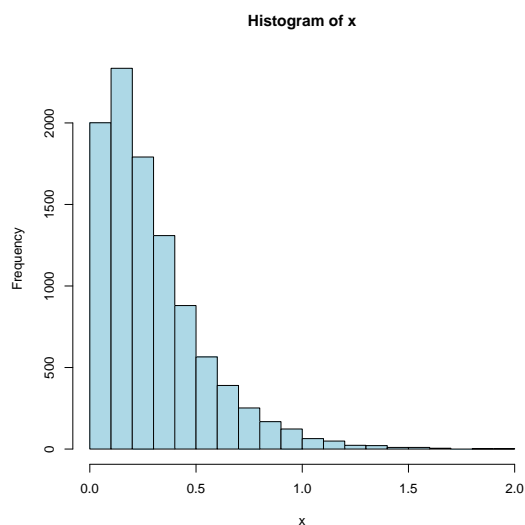
Fractions and counts

## Why is a histogram not a bar graph?

- Frequencies are represented by **area**, not height.
- There is no space between the bars.
- The horizontal axis represents a numerical quantity, with an inherent order.

## Interpreting histograms

- Describe the overall pattern and any significant deviations from that pattern.
- **Shape:** Is the distribution (approximately) symmetric or skewed?



This distribution is skewed **right** because it has a long right-hand tail.

- **Center:** Where is the “middle” of the distribution?
- **Spread:** What are the smallest and largest values?
- **Outliers:** Are there any observations that lie outside the overall pattern? They could be unusual observations, or they could be mistakes. **Check them!**