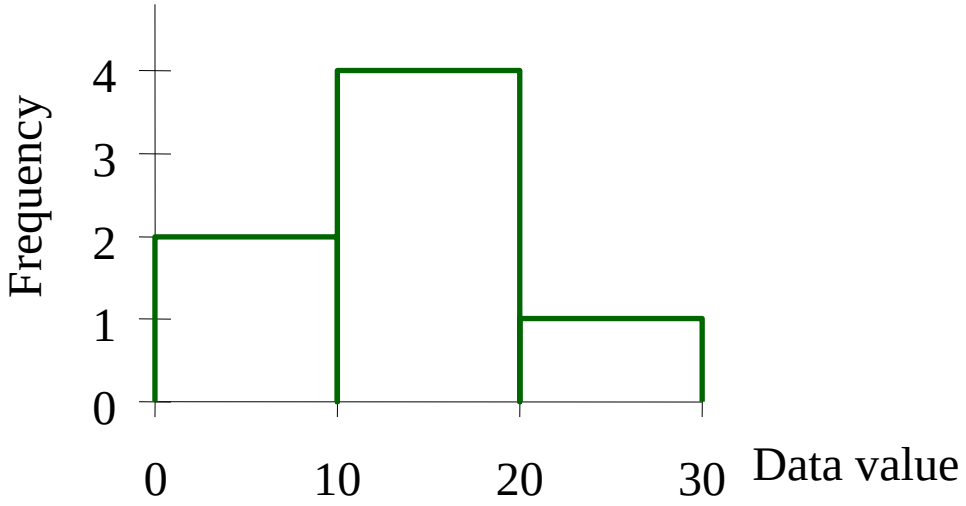
# Chapter 3

## Histograms: Looking at the Distribution of the Data

# Histogram

- A Picture of a list of numbers

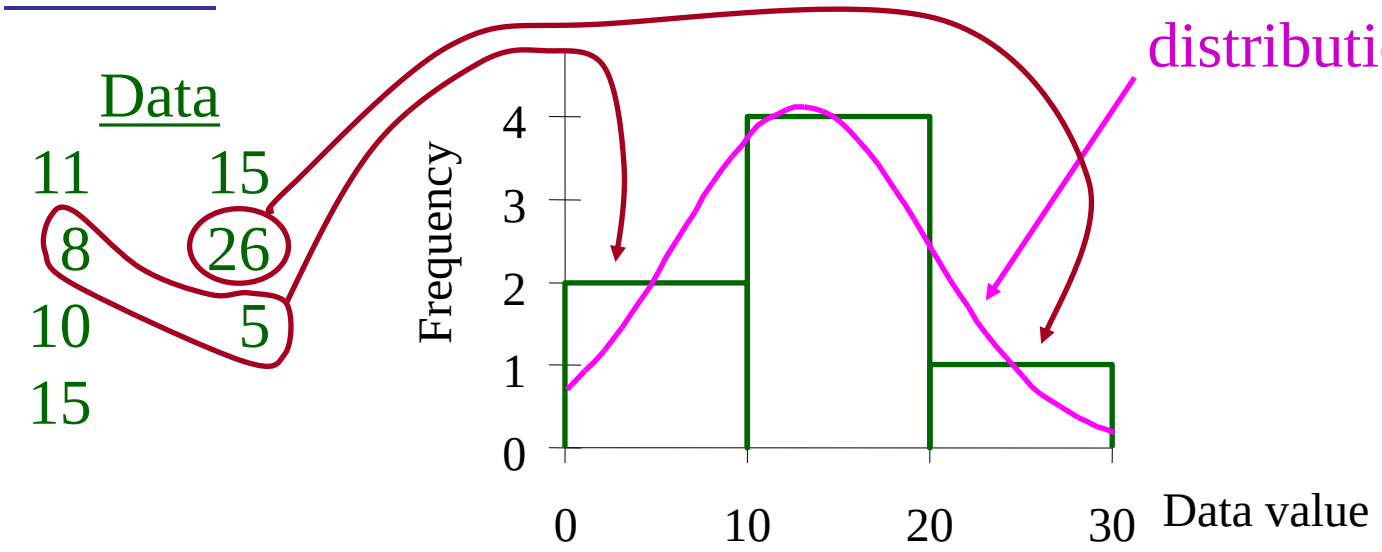
| Data | |
|---|---|
| 11 | 15 |
| 8 | 26 |
| 10 | 5 |
| 15 | |




- BARS ARE HIGH when many elementary units fall within this range
- Shows typical value (center), dispersion (variability), distribution shape, outliers (if any)

# Histogram

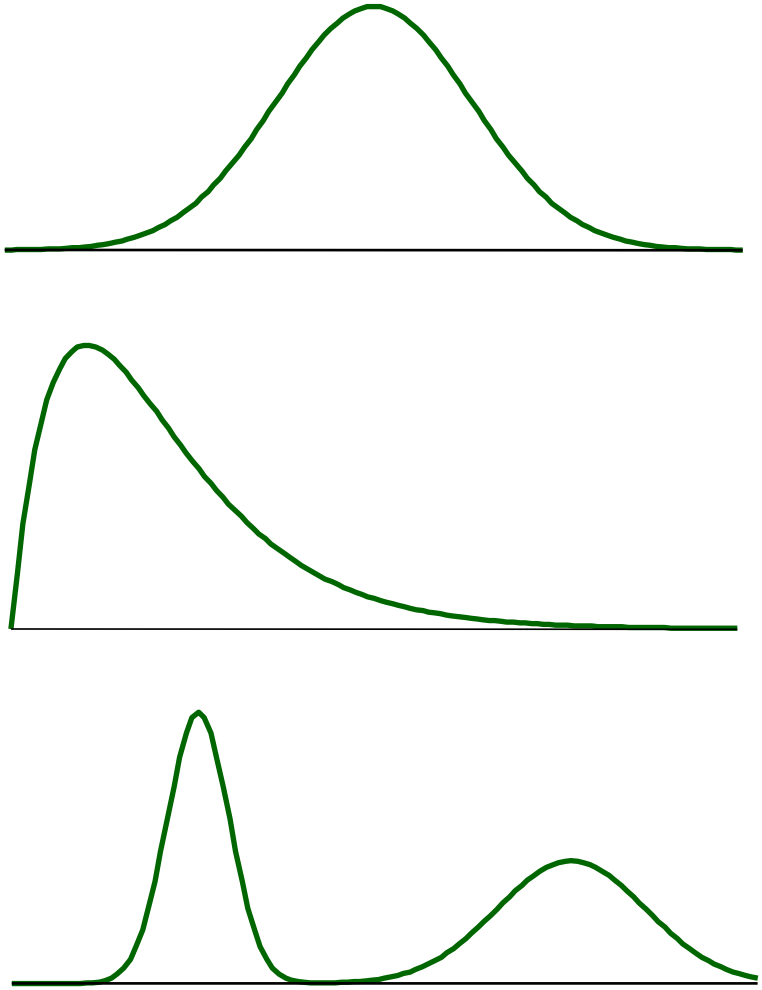
- A Picture of a list of numbers




- BARS ARE HIGH when many elementary units fall within this range
- Shows typical value (center), dispersion (variability), distribution shape, outliers (if any)

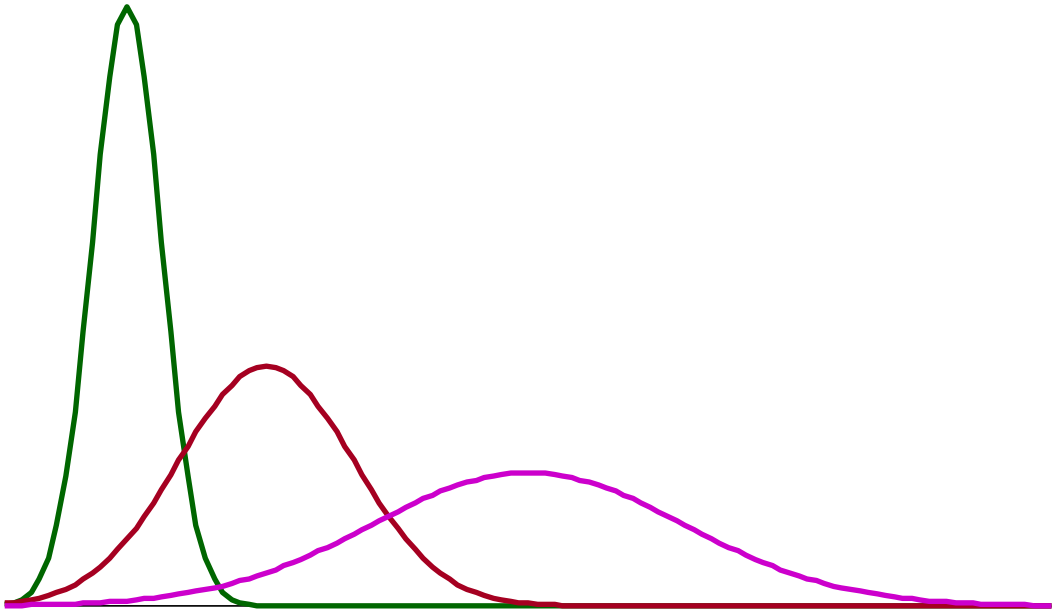
# Distribution Shapes (Ideal)

- Normal
  - Symmetric
  - Bell-Shaped
- Skewed
  - Not symmetric
  - Can cause trouble
  - Transform? Logarithm?
- Bimodal
  - Two clear groups
  - Find out why!
  - Analyze separately?



# Idealized Normal Distributions

- Can shift center, width (diversity) of distribution
- In idealized form, without the randomness of data



# Data from a Normal Distribution

- All are sampled from the same idealized normal distribution. Note the random differences.

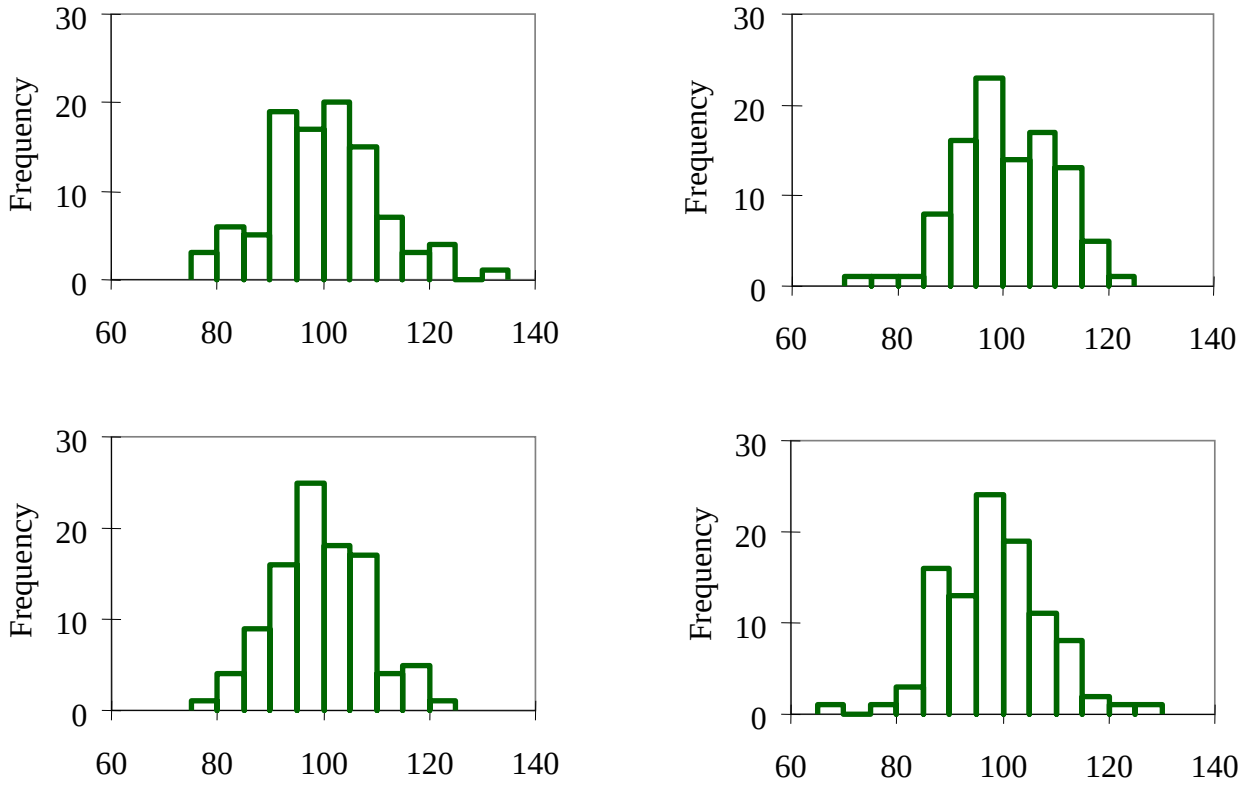
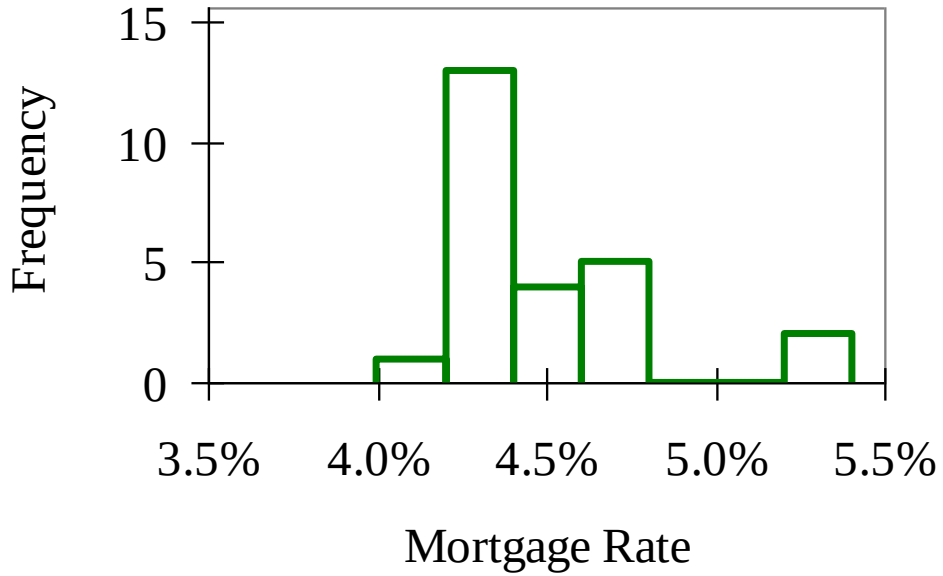


Fig 3.2.1

# Example: Mortgage Interest Rates

- Values from about 4.0% to 5.4%
- Typical: from about 4.2% to 4.8%
- Diversity among institutions
- Special feature: gap from 4.8% to 5.2%



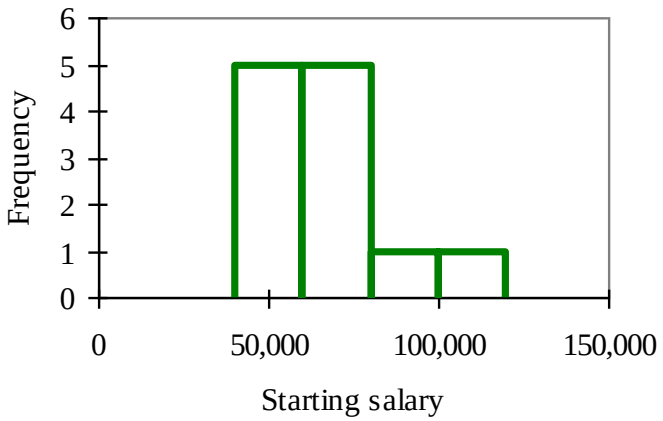
# Histogram and Bar Chart

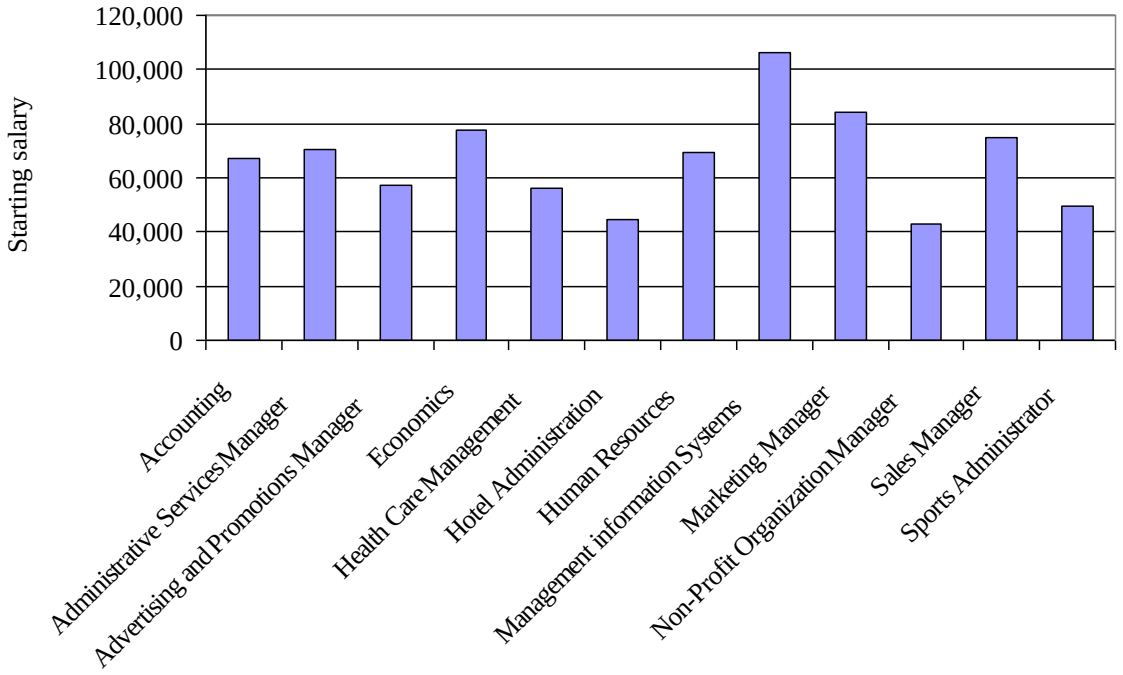
- Histogram is a bar chart of the frequencies of the data
  - Histogram: bar height represents number of cases within the range
  - Ordinary bar chart: bar height represents data value for just one case
- Histogram shows overall distribution
  - Histogram: the "big picture" of patterns in the data
  - Ordinary bar chart: often  too much detail (each individual case)

Fig 3.2.2-3

# Histogram and Bar Chart for Salaries

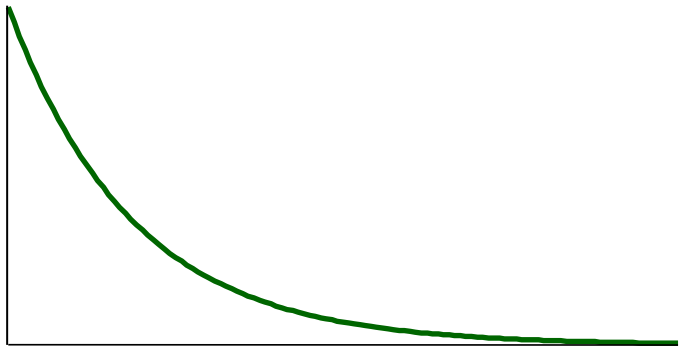
- Histogram shows patterns in the *frequencies*
- Ordinary Bar chart shows *all cases individually*
- For large data sets, histogram is much more useful



# Idealized Skewed Distributions

- Not symmetric
- Various shapes are possible
- In idealized form, without the randomness of data



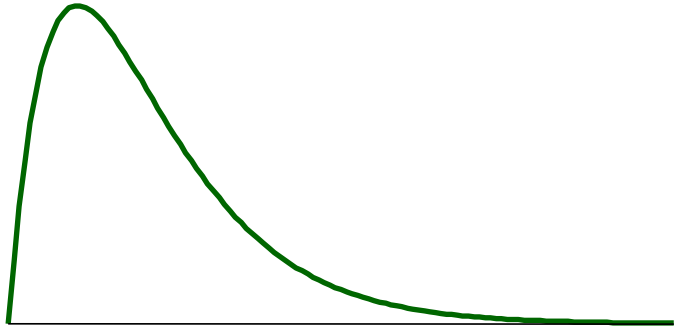


Fig 3.4.2

# Example: Bank Deposits

- Most banks are smaller: tall bars at the left
- A few banks are larger (to the right)
- A skewed distribution

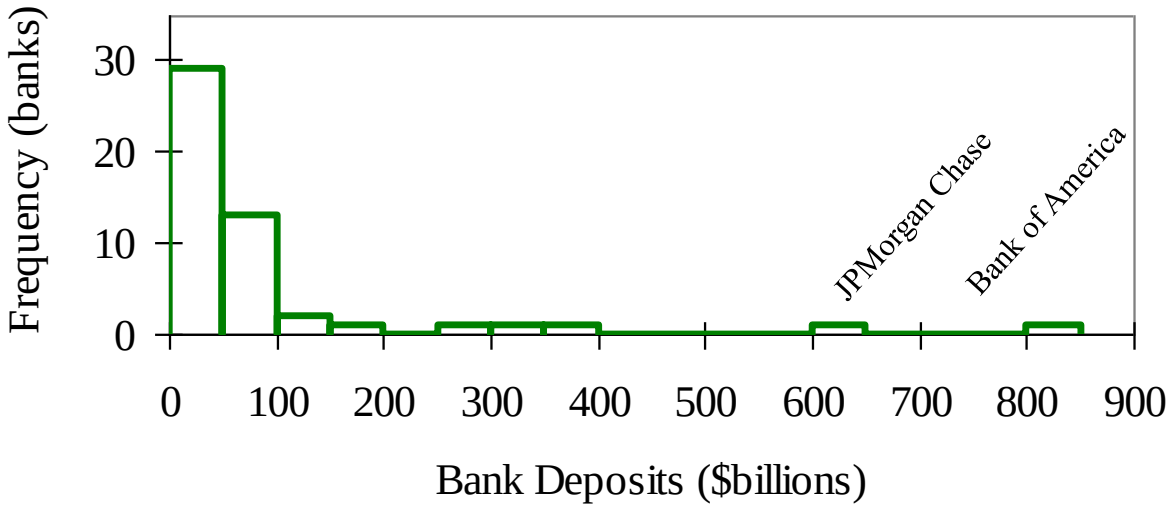
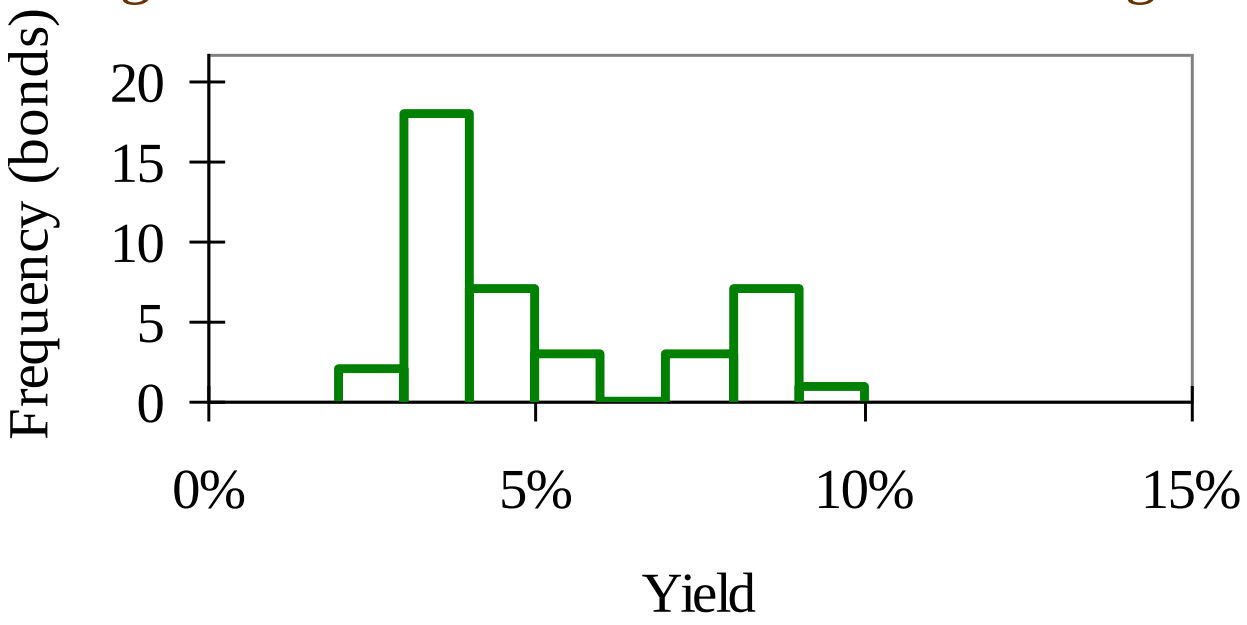


Fig 3.5.1

# Bimodal Distribution

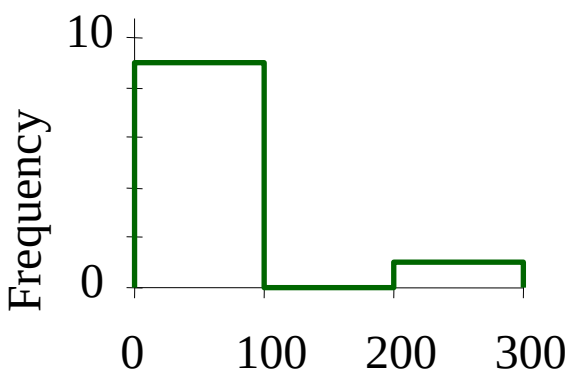
- Two distinct groups in the data (ask "why?")
- Example: Corporate Bonds rated AA and B
  - Low-risk bonds have a lower yield
  - High-risk bonds entice investors with a higher yield

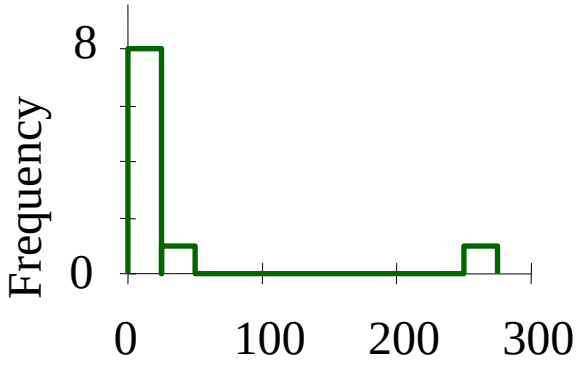


# Outlier

- A data value very different from the others
- Difficult to see distribution of most of the data, even after changing histogram scale

| Defects | |
|---|---|
| 11 | 19 |
| 23 | 15 |
| 18 | 19 |
| 13 | 268 |
| 25 | 9 |



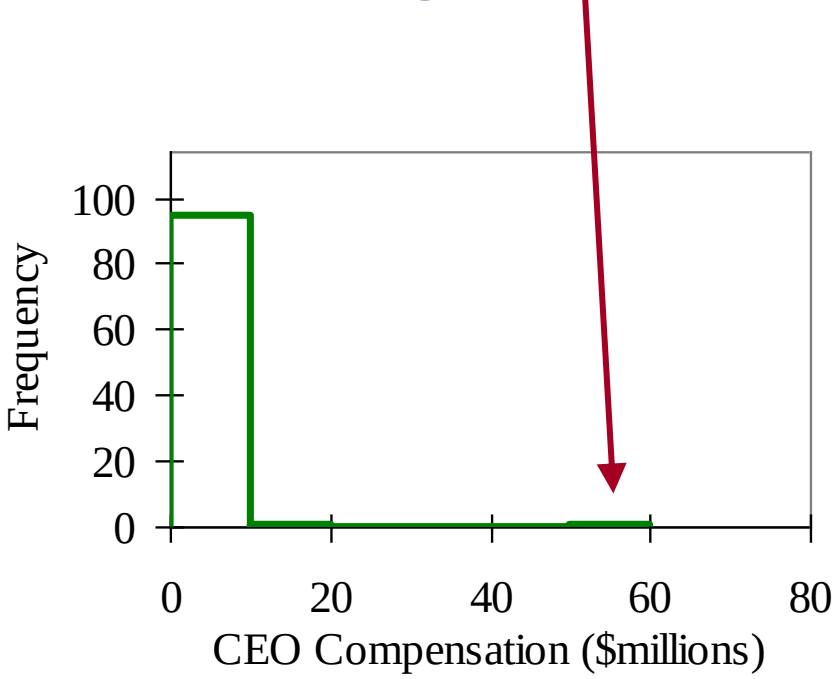
# Outlier: What to Do?

- Note the outlier. If error, then fix it
- (Perhaps) analyze with and without outlier(s)
  - If similar answers, then no problem
- OK to omit outlier(s) **IF** not part of situation under study
  - e.g., Lab analysis, dropped test tube
    - OK to omit, if studying normal operation, not laboratory accidents
  - e.g., Statistical audit, "special occurrence" error
    - Use care. Such an error in a sample may represent other "explainable" errors in accounts that were not examined

# Example: Software CEO Compensation

Fig 3.6.1, 4

- One CEO (Ellison of Oracle) made $56.81 million
- Removing this outlier, we can see more detail



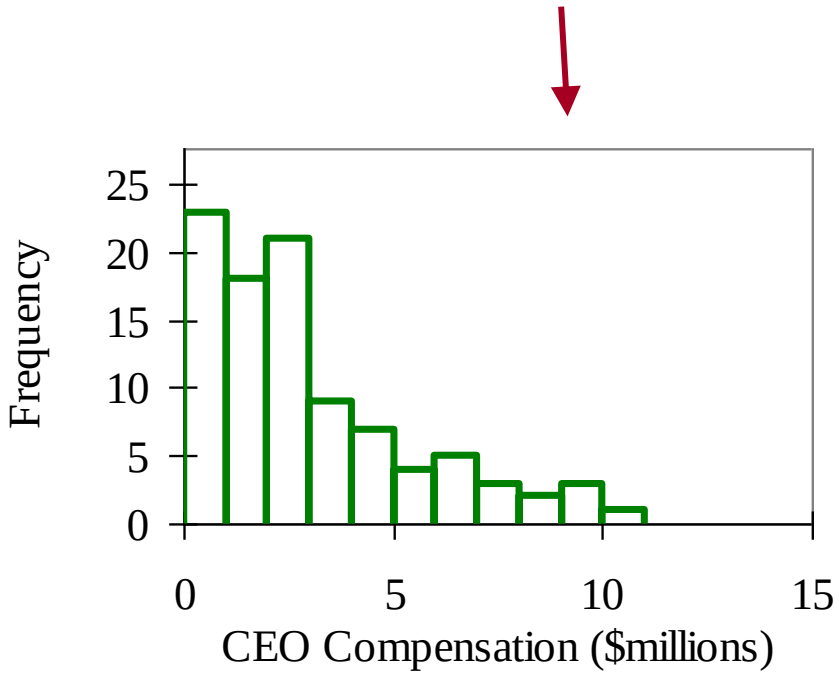



Fig 3.7.1

# Data Mining Promotions Received

- Number of promotions received by 20,000 people in the donations database

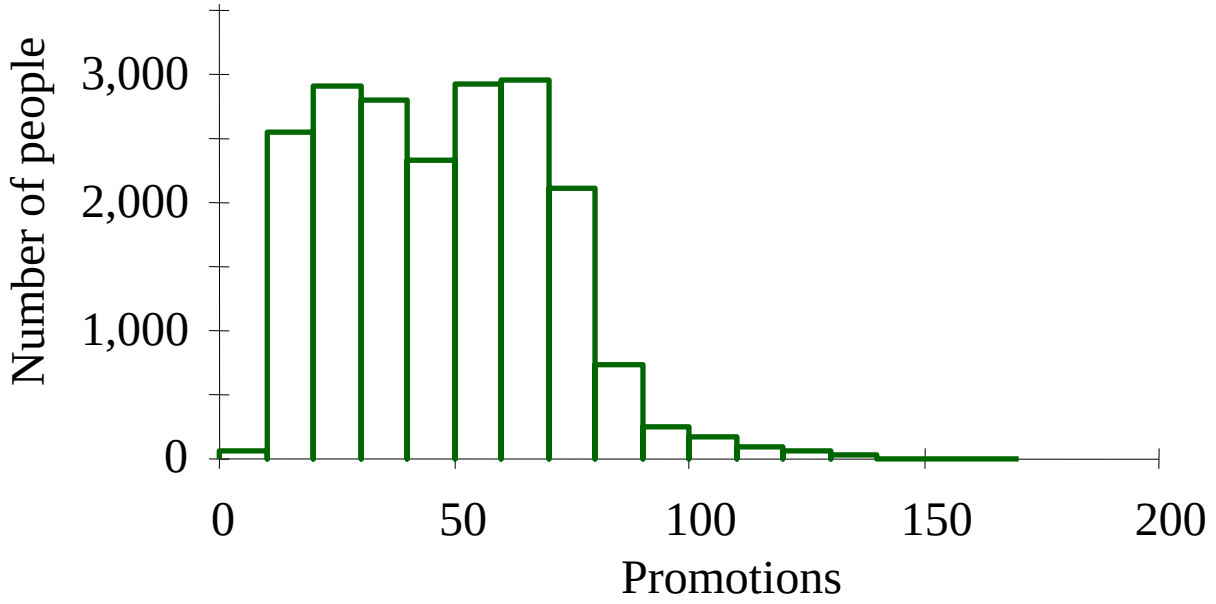


Fig 3.7.2

# More Detail in Promotions

- Reduce bar width from 10 to 1 promotion
- With large data set, can see interesting structure
  - such as the peak at about 15 promotions

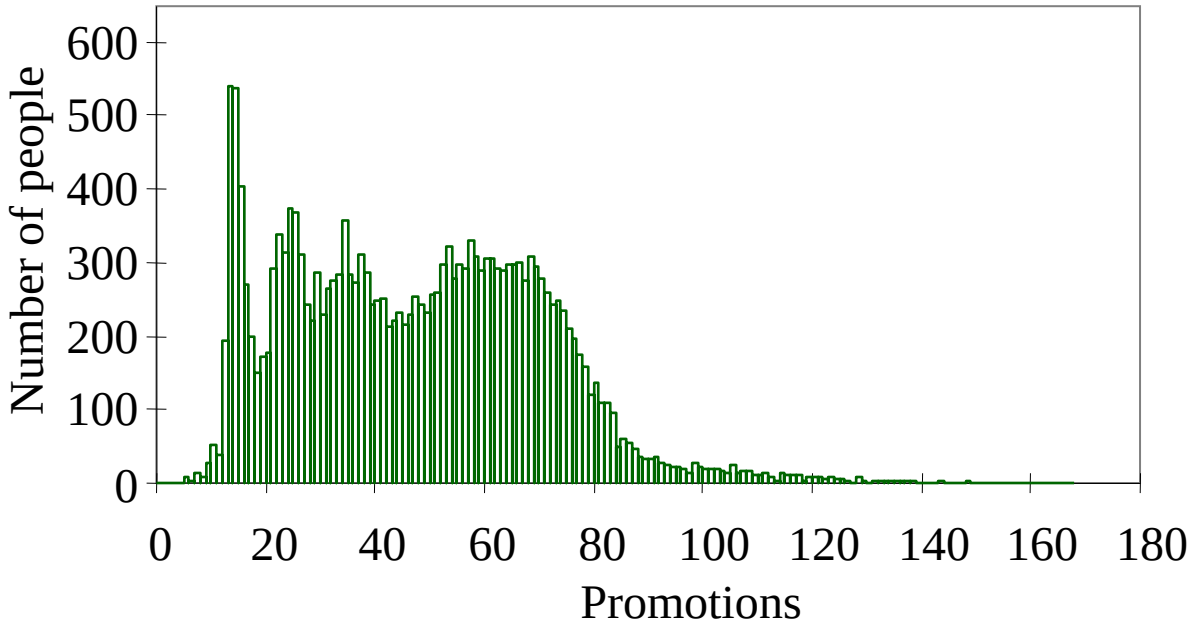


Fig 3.7.3

# Data Mining Donations

- Size of donation received in response to mailing
- Note: many donations of $0 among these 20,000
  - Difficult to see anything else! (six donated $100)

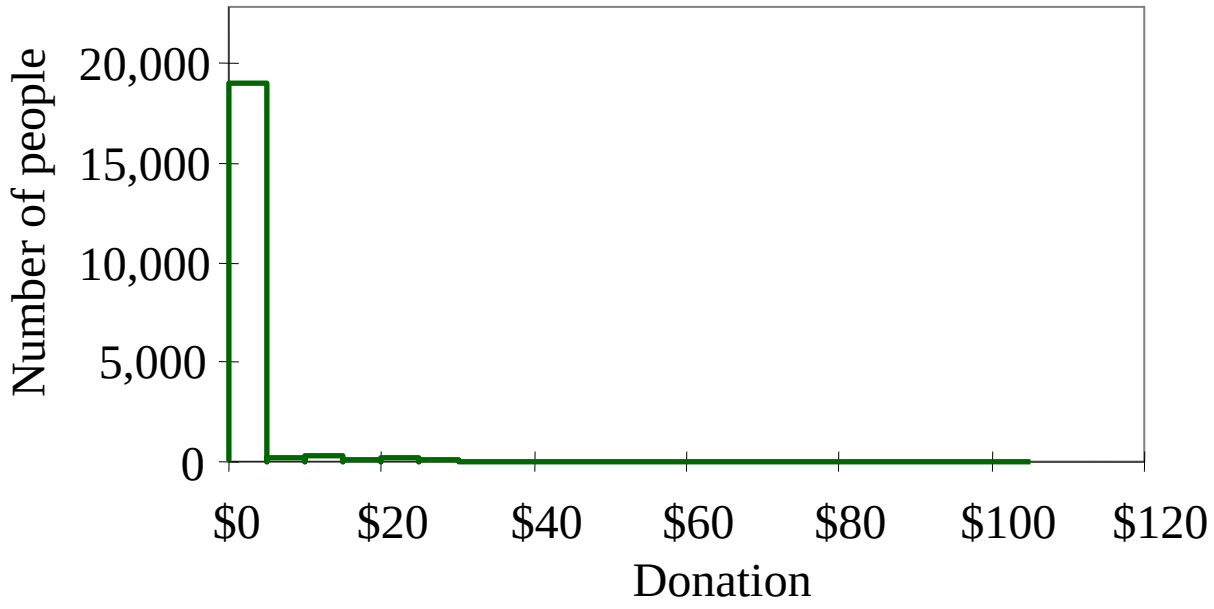


Fig 3.7.4

# More Detail in Donations

- Keep only the 989 who donated (eliminate $0)
  - to see detail among those who made a gift
- Can now see the distribution of the gift amounts

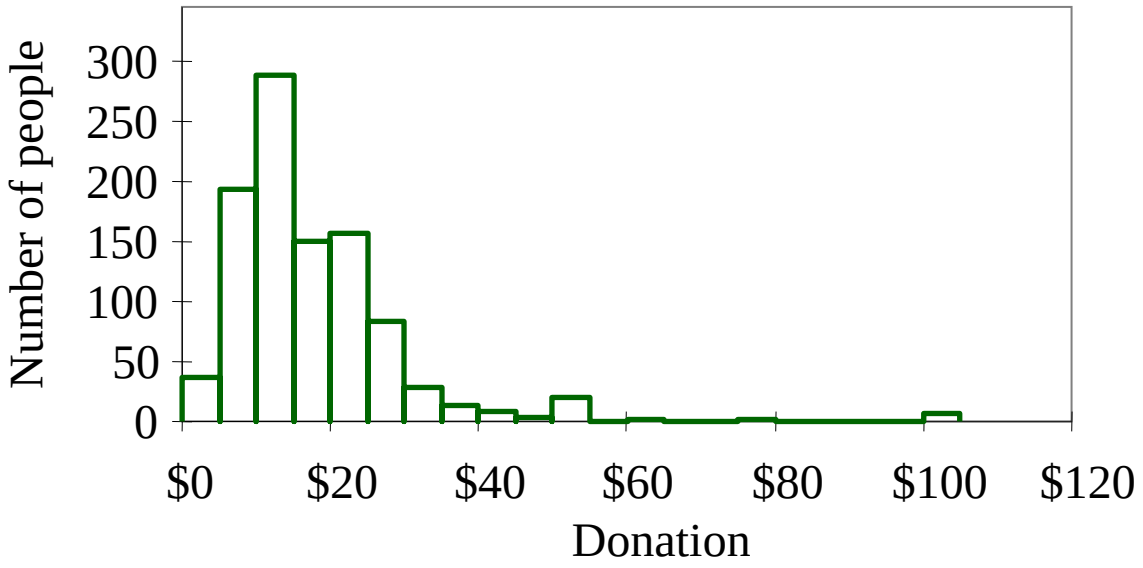


Fig 3.7.5

# Even More Detail in Donations

- With so much data (989 people)
  - we can use smaller bars to see more details
- Note the "spikes" at $5, 10, 15, 20, 25, and 50