Statistical Foundation for Business Analytics

Sumeet Gupta

Professor (Information Technology and Systems)

Indian Institute of Management Raipur

Outline

- Variation and Variance
- Summarizing Variance
- Statistical Inference: Random Sampling
- Confidence Interval and Hypothesis Testing
- Hypothesis Testing for Ordinal Data
- Hypothesis Testing for Nominal Data

Variation and Variance

Principle of Variation

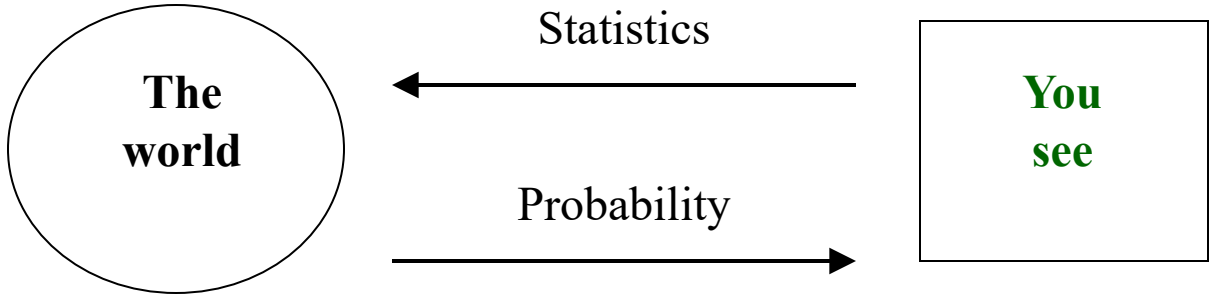
- Constant: Whose value does not change
  - E.g. Gravitational Constant for earth
- But most observations in life vary
  - E.g. Rising time of the sun; Amount we eat everyday; Sales; Production figures
- Can we capture this variation?
  - If the variation is small, our predictions can be fairly accurate
    - E.g. Rising time of the sun
  - But if the variation is large, our predictions remain inaccurate
    - E.g. Occurrence of an Earthquake

Data Science

- Capturing this variation and making predictions
  - Data ➔ Estimate the Unknown
- Statistics: Art and Science of Collecting and Understanding DATA
  - DATA = Recorded Information (Sales, Productivity, Quality, Costs, Return, …)
- Data Mining: Search for patterns in large data sets
  - Statistics: Prediction, Classification and Clustering
  - Computer Science: efficient algorithms (instructions) for collecting, maintaining, organizing, analyzing data
  - Optimization: calculations to achieve a goal

Probability

- "Inverse" of statistics



  - **Statistics**: generalizes from data to the world
  - **Probability**: "What if …" Assuming you know how the world works, what data are you likely to see?
- Examples of probability:
  - Flip coin, stock market, future sales, IRS audit, …
- Foundation for statistical inference

Types of Variables

- A Variable
  - The type of measurement being done
    - e.g., Sales volume, Cost, Productivity, Number of defects, …
- Types
  - Quantitative variables (Ratio Vs Interval)
  - Qualitative (Nominal Vs Ordinal)
- Collection
  - Cross-Sectional (No meaningful sequence)
  - Over a period of time (Meaningful sequence – Time Series)

How Many Variables?

- Univariate data set: One variable measured for each elementary unit
  - e.g., Sales for the top 30 computer companies.
  - Can do: Typical summary, diversity, special features
- Bivariate data set: Two variables
  - e.g., Sales and # Employees for top 30 computer firms
  - Can also do: relationship, prediction
- Multivariate data set: Three or more variables
  - e.g., Sales, # Employees, Inventories, Profits, …
  - Can also do: predict one from all other variables

Visualizing and Summarizing Variance

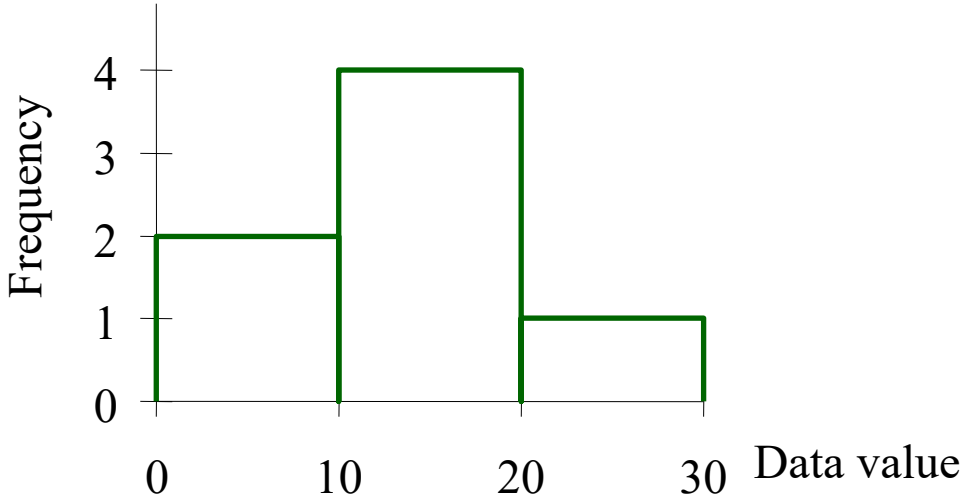
Visualizing and Summarizing Variance

- Quantitative Variables
  - Visual Representation: Histograms and Boxplots
  - Central Tendency: Mean, Median and Mode
  - Dispersion: Std Deviation, Variance, Range, Coefficient of variation
- Qualitative Variables
  - Visual Representation: Bar Charts, Pie Charts
  - Central Tendency: Median and Mode
  - Dispersion: No of Categories
- Time Series Data
  - Visual Representation: Time Series Chart
  - Descriptors: Trend, Seasonality and Cyclicity

Histograms

- A Picture of a list of numbers

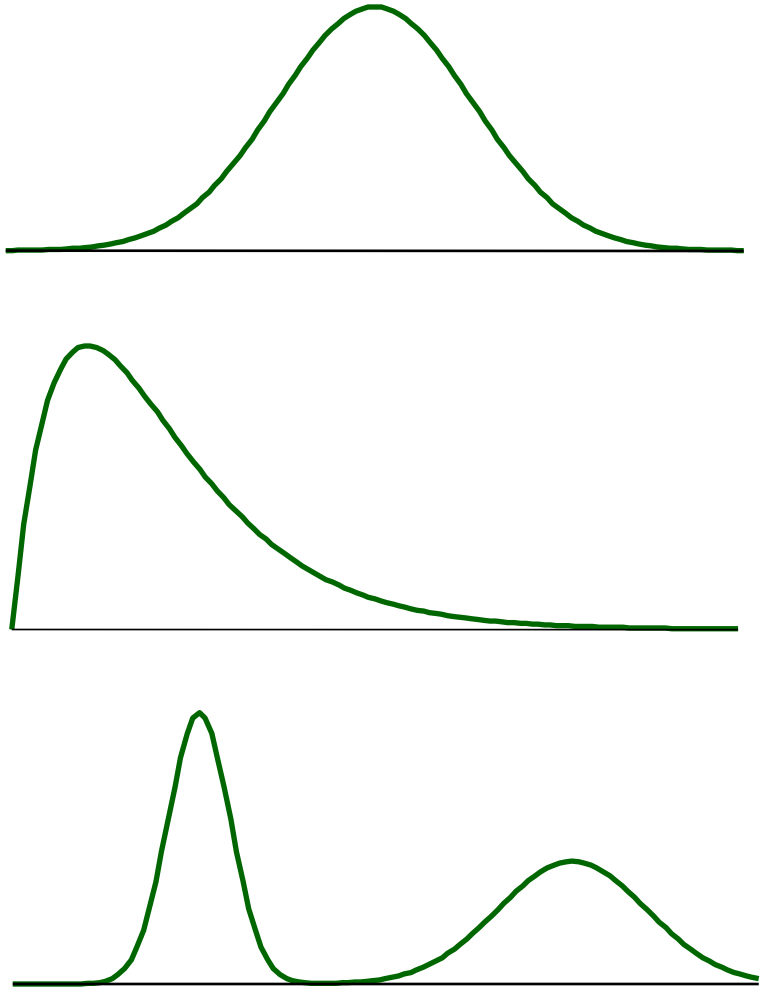
Data	
11	15
8	26
10	5
15	



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

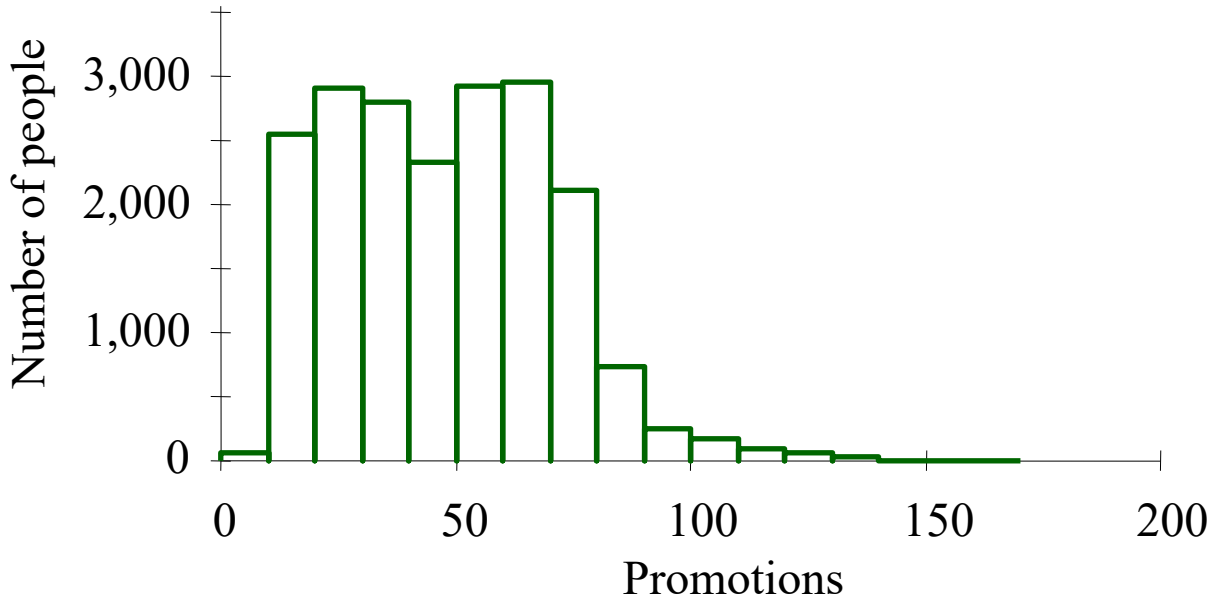
Histogram: Ideal Shapes

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



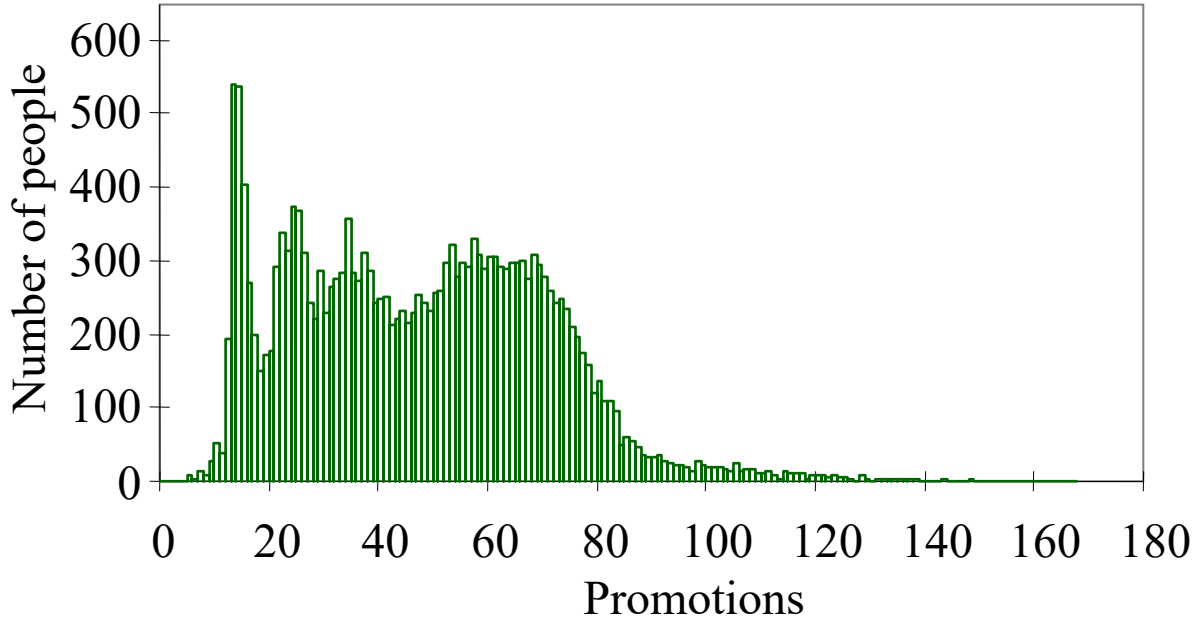
Histogram: Example

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



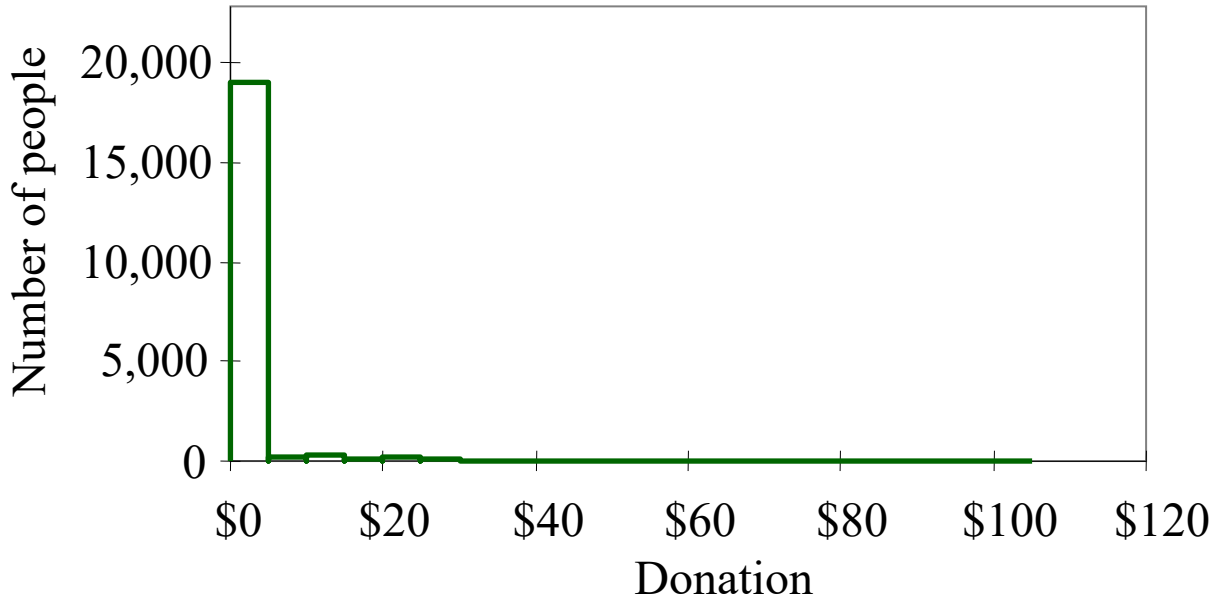
Histogram: Example

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



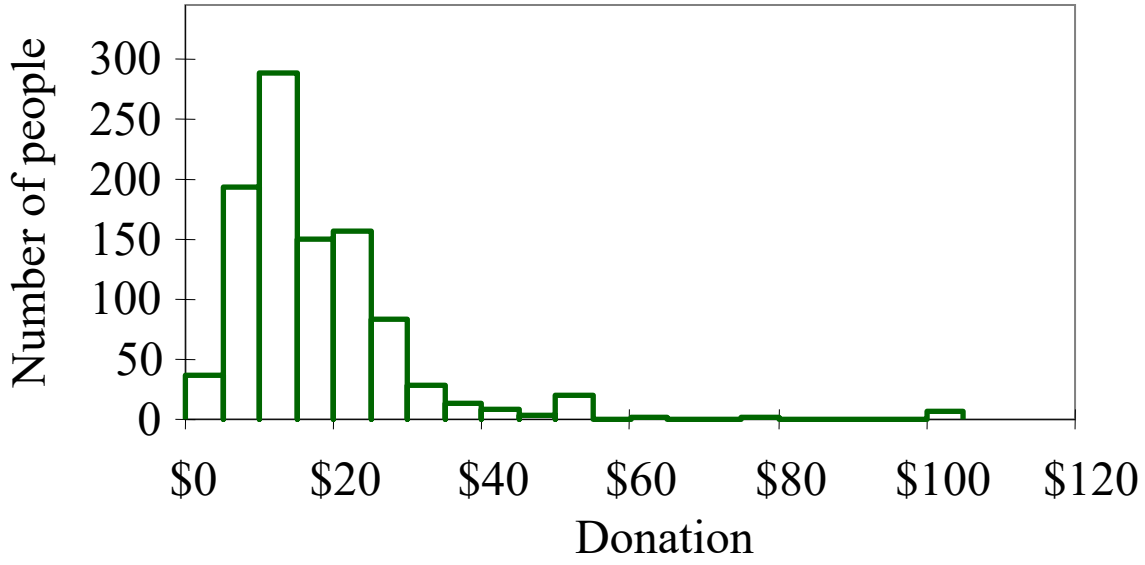
Histogram: Example

- **Donations Dataset**
- 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)



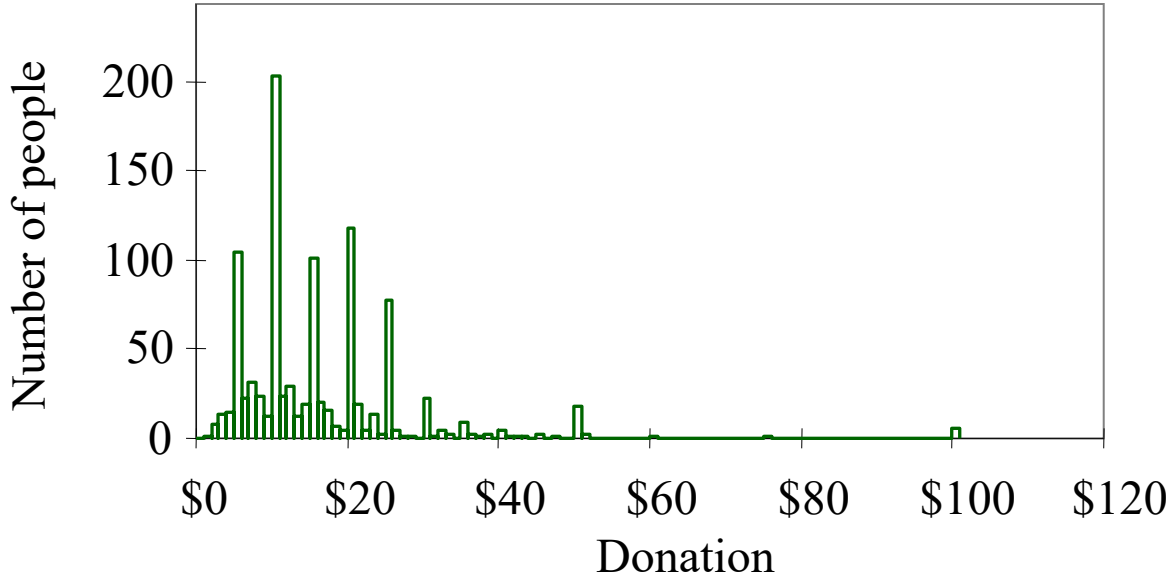
Histogram: Example

- **Donations Dataset**
- 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



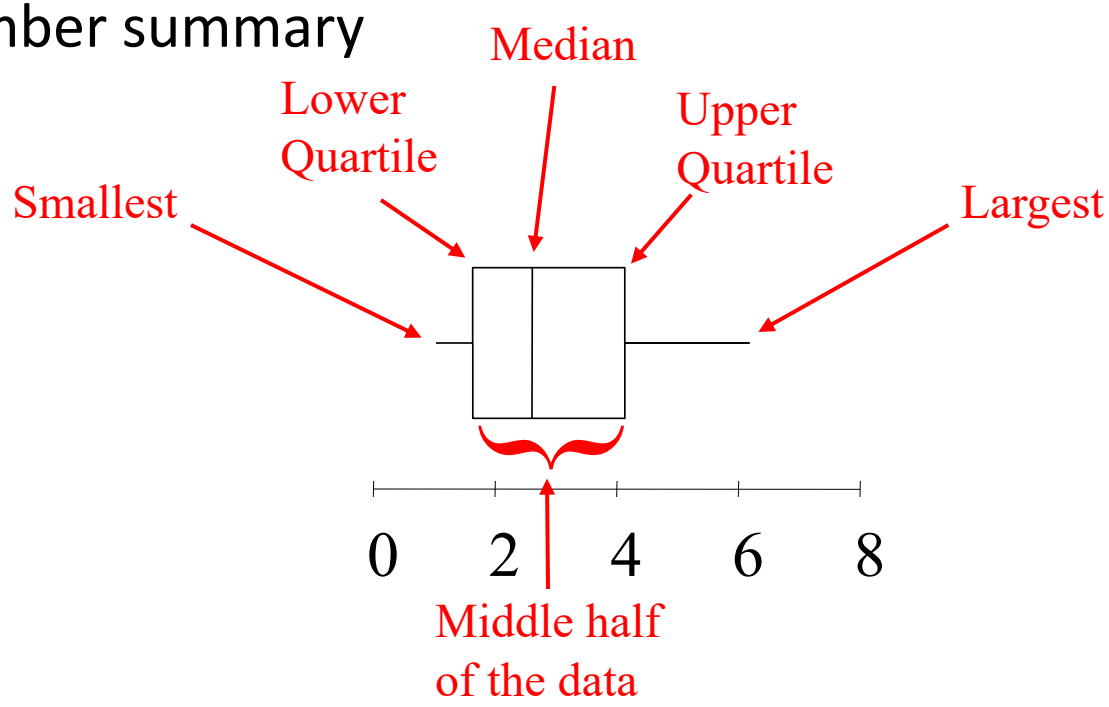
Histogram: Example

- **Donations Dataset**
- 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



Box Plot

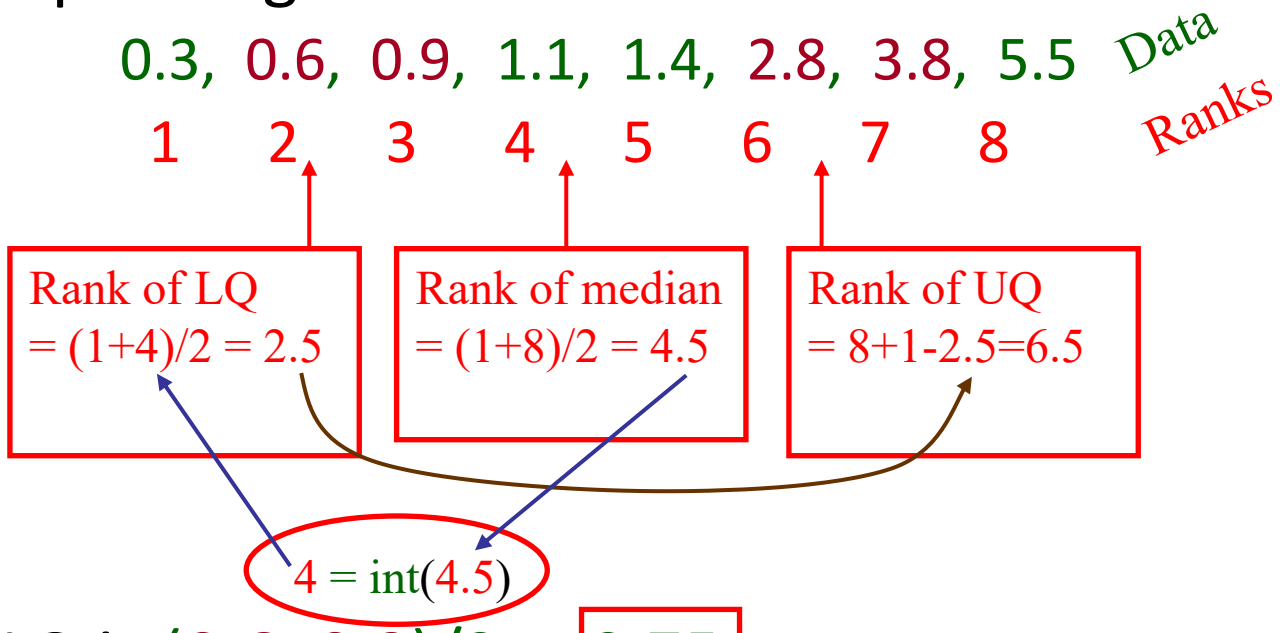
- Displays five-number summary



- Less detail than histogram
  - Easier to compare many groups

Box Plot: Example

- Spending rank ordered from smallest to largest



- LQ is (0.6+0.9)/2 = 0.75
- UQ is (2.8+3.8)/2 = 3.3

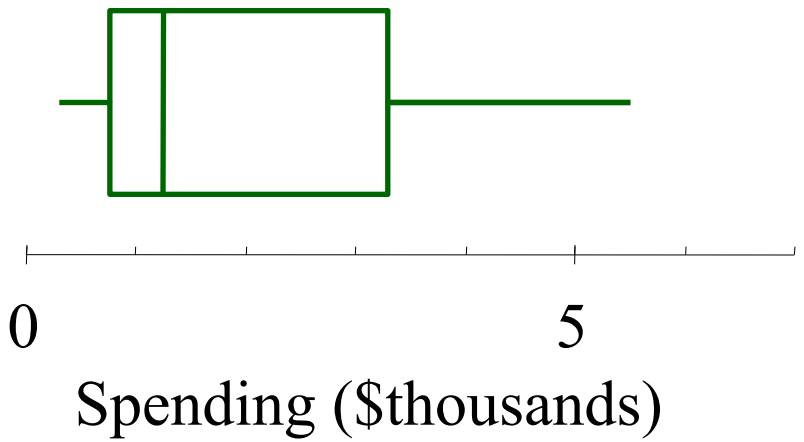
Box Plot: Example

- Five-number summary

  0.3, 0.75, 1.25, 3.3, 5.5

  Smallest, LQ, Median, UQ, Largest

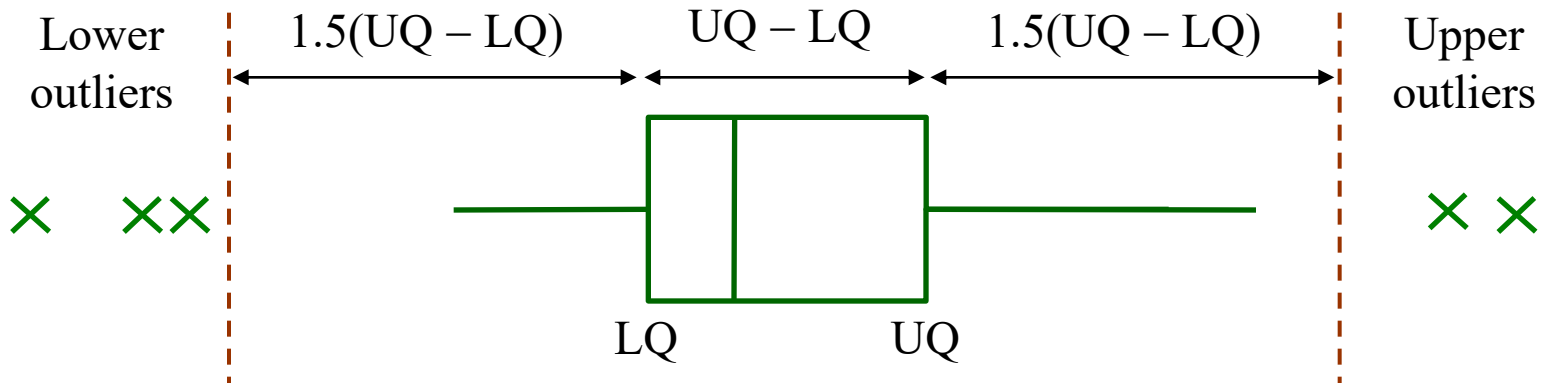
- Box plot



  - Shows some skewness (lack of symmetry)

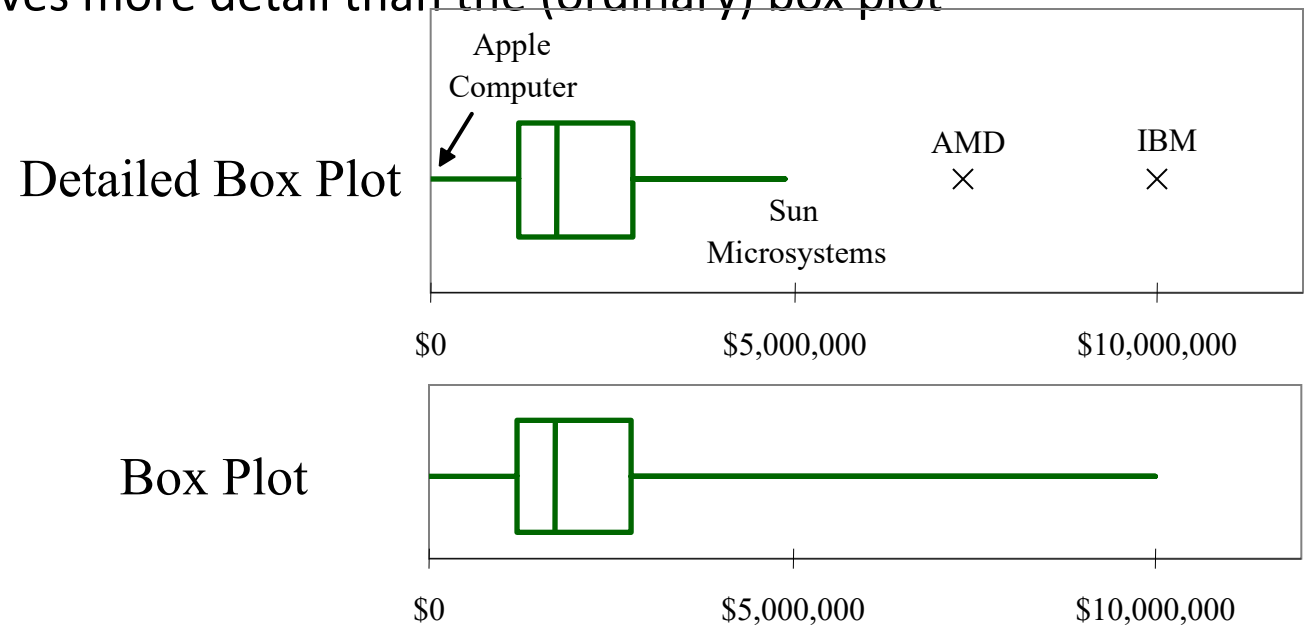
Box Plot: Identifying Outliers

- Outliers are defined as observations, if any, either:
  - More than UQ + 1.5 (UQ – LQ), or
  - Less than LQ – 1.5 (UQ – LQ)
- Outliers are far from the center of the distribution
  - and may be interesting as special cases



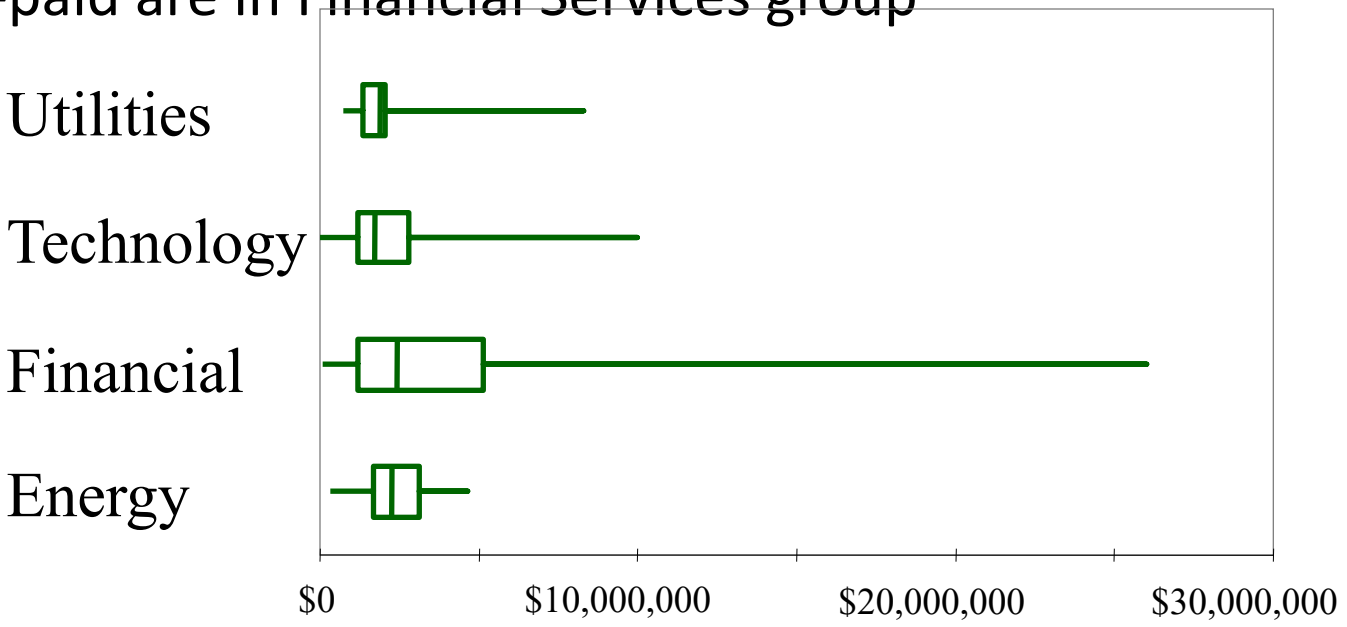
Box Plot: Example - CEO Compensation

- CEO compensation in technology companies
  - Detailed box plot identifies outliers
    - and identifies the most extreme non-outliers,
    - gives more detail than the (ordinary) box plot



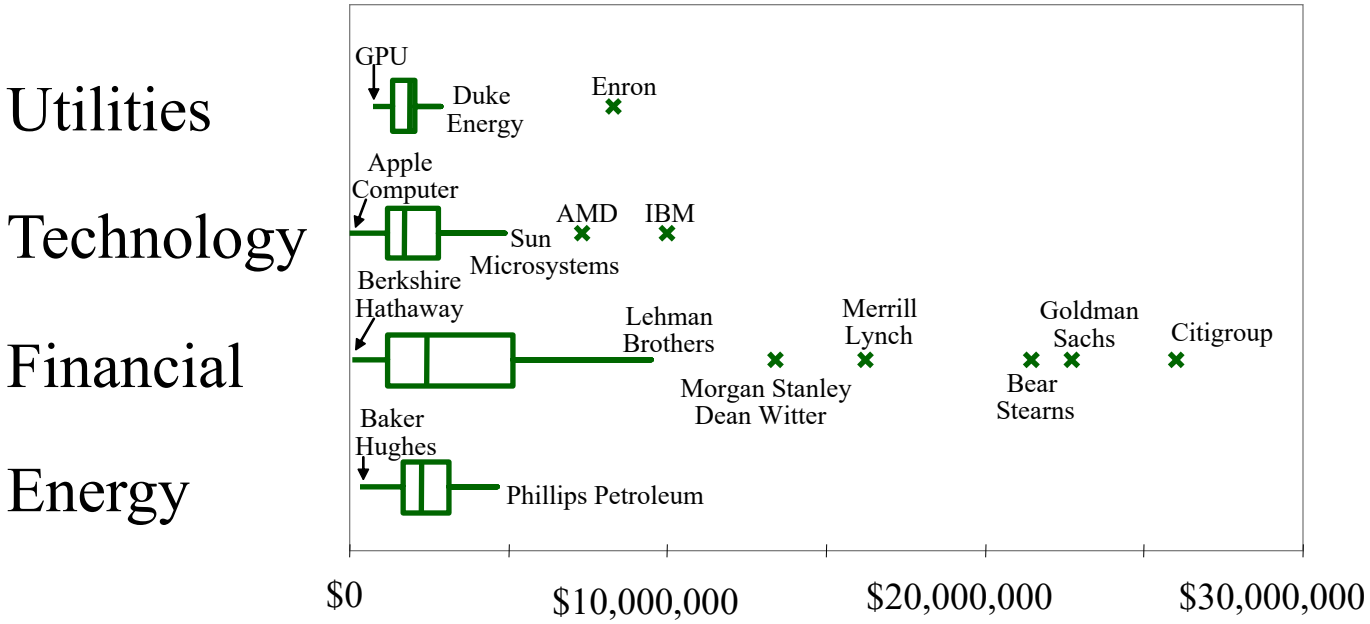
Box Plot: Example - CEO Compensation

- Box plots to compare firms within industry groups
  - Utilities group generally shows lower compensation
  - Highest-paid are in Financial Services group



Box Plot: Example - CEO Compensation

- Detailed box plots (with outliers and most extreme non-outliers named)



Central Tendency: Average or Mean

- Add the data, divide by *n* or *N* (the number of elementary units)

"X-bar"

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Sample average

"mu"

$$\mu = \frac{X_1 + X_2 + \ldots + X_N}{N}$$

Population average

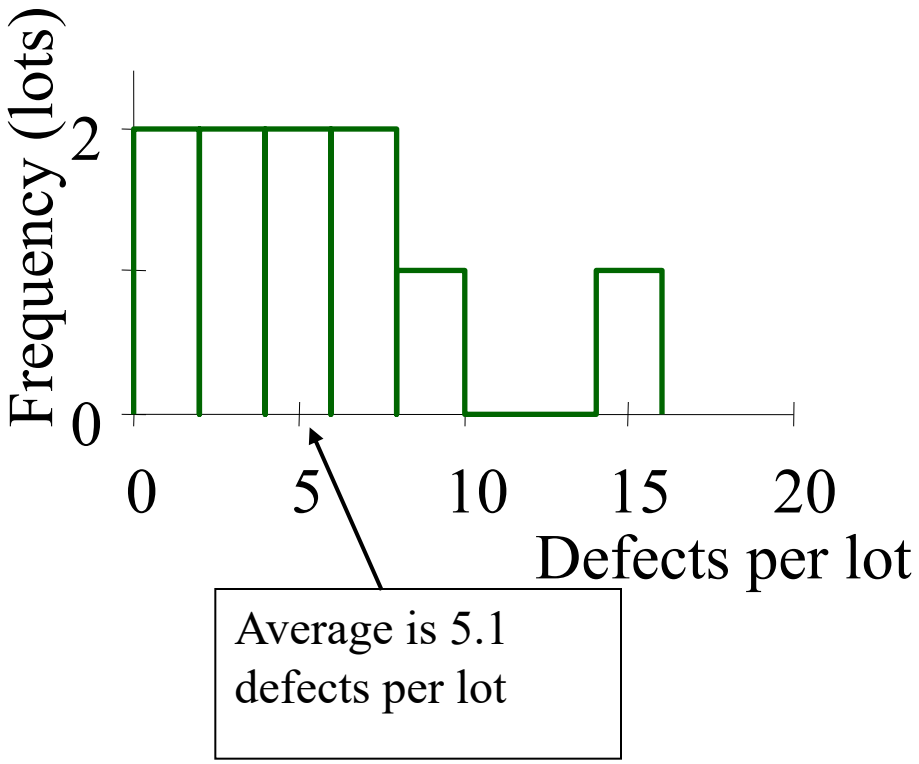
- Divides total equally. The only such summary
- A representative, central number (if data set is approximately normal)
- Summation notation
  - Σ is capital Greek sigma

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Example: Number of Defects

- Defects measured for each of 10 production lots

  4, 1, 3, 7, 3, 0, 7, 14, 5, 9



Central Tendency: Median

- A representative, central number
  - If data set *has* a center
- Less sensitive to outliers than the average
- For skewed data, represents the "typical case" better than the average does
  - e.g., incomes
    - Average income for a country equally divides the total, which may include some very high incomes
    - Median income chooses the middle person (half earn less, half earn more), giving less influence to high incomes (if any)

Central Tendency: Median

- Also summarizes the data
- The **middle** one
  - Put data in order
  - Pick middle one (or average middle two if *n* is even)
  - Median (9, 4, 5) = Median(4, 5, 9) = 5
  - Median (9, 4, 5, 7) = Median (4, 5, 7, 9) = (5+7)/2 = 6
- Rank of the median is $(1+n)/2$
  - If $n=3$, rank is $(1+3)/2 = 2$
  - If $n=4$, rank is $(1+4)/2 = 2.5$ (so average 2nd and 3rd)
  - If $n=262$, rank is $(1+262)/2 = 131.5$

Central Tendency: Median Example

- Customers plan to spend ($thousands)

  3.8, 1.4, 0.3, 0.6, 2.8, 5.5, 0.9, 1.1

- Rank ordered from smallest to largest

  0.3, 0.6, 0.9, 1.1, 1.4, 2.8, 3.8, 5.5

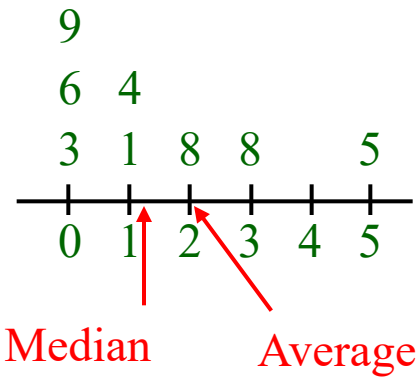
  1 2 3 4 5 6 7 8

Data

Ranks

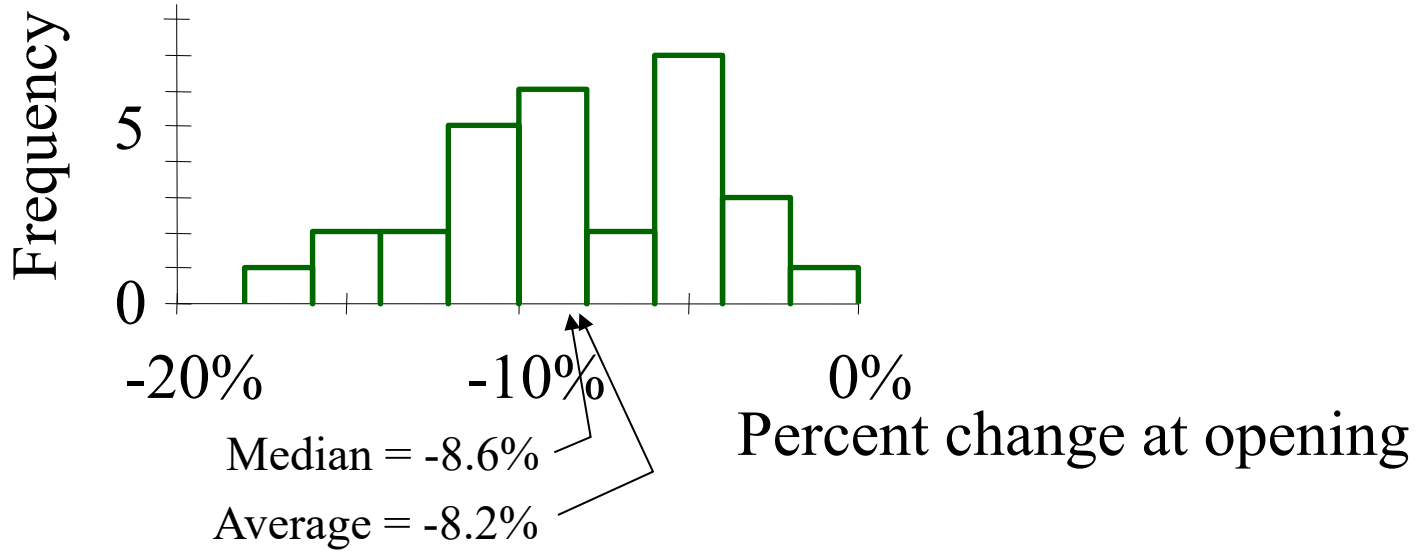
Rank of median
= (1+8)/2 = 4.5

- Median is (1.1+1.4)/2 = 1.25
  - Smaller than the average, 2.05
    - Due to slight skewness?



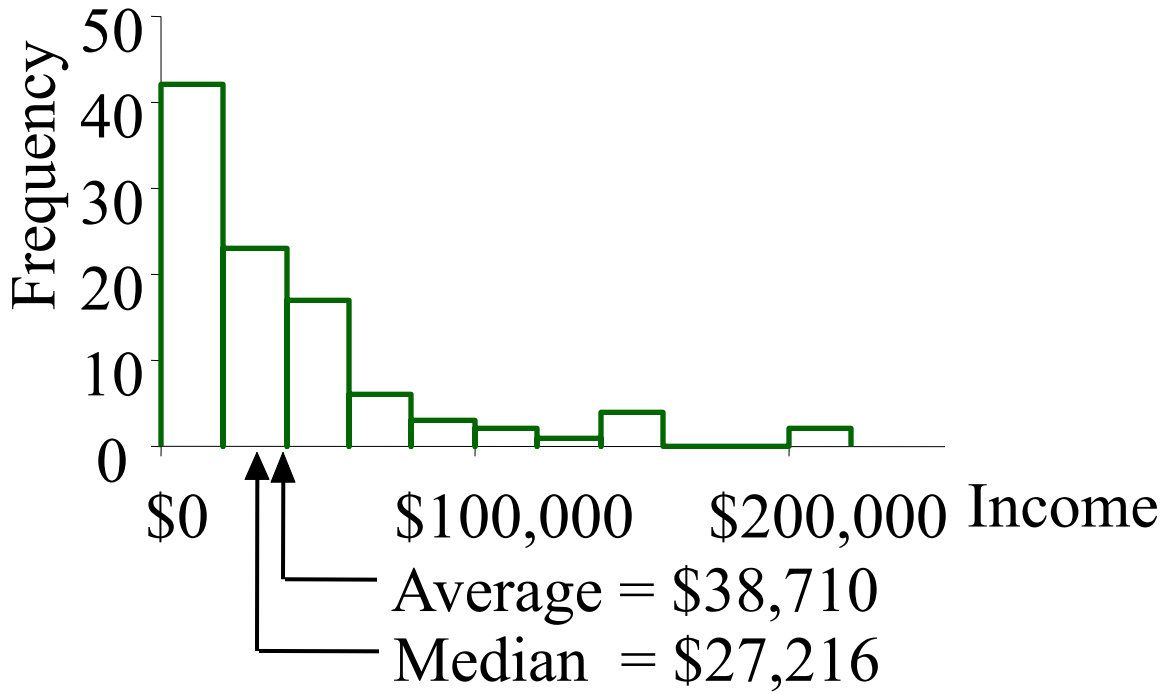
Central Tendency: Median Example

- **Example: The Crash of 1987**
- Dow-Jones Industrials, stock-price changes as each stock began trading that fateful morning
- Fairly normal
- Mean and median are similar



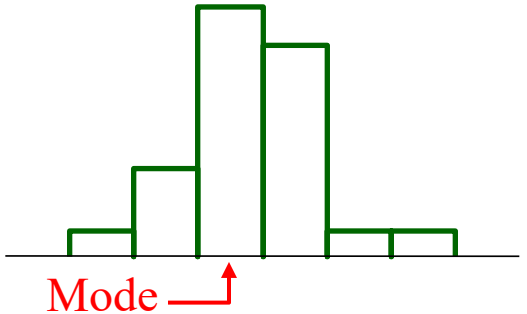
Central Tendency: Median Example

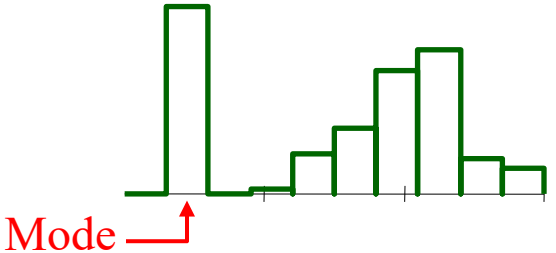
- **Personal income of 100 people**
- Average is higher than median due to skewness



Central Tendency: Mode

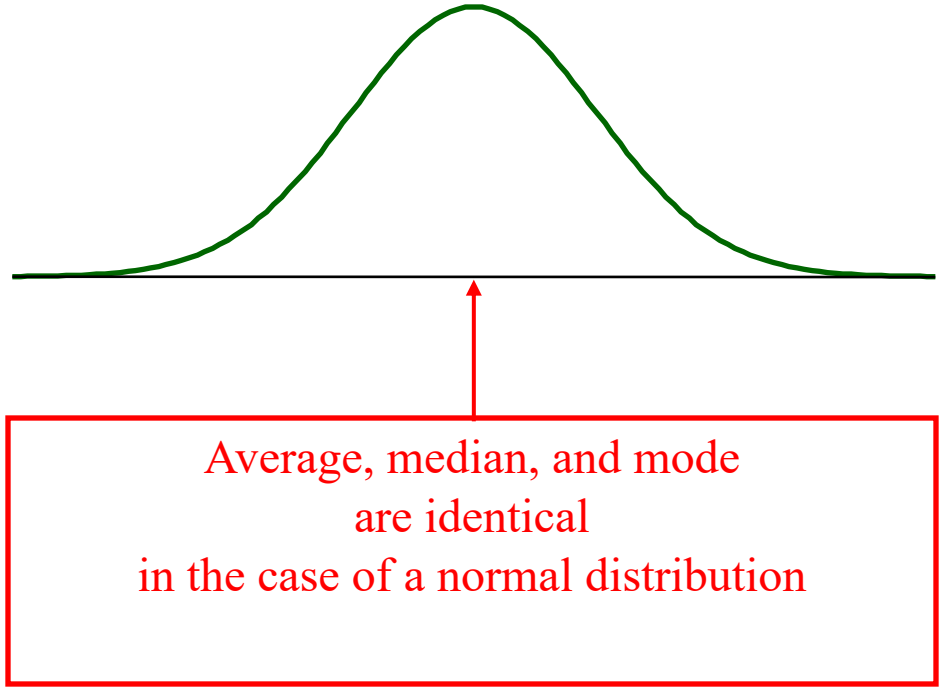
- Also summarizes the data
- Most common data value
  - Middle of tallest histogram bar
- Problems:
  - Depends on how you draw histogram (bin width)
  - Might be more than one mode (two tallest bars)
- Good if most data values are “correct”
- Good for nominal data (e.g., elections)





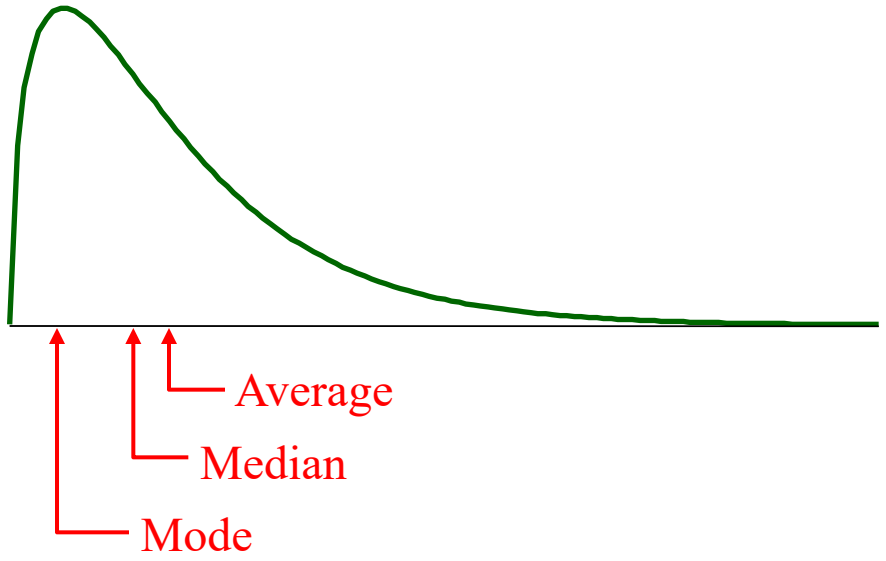
Normal Distribution

- Average, median, and mode are **identical**
  - If the data come from a normal distribution



Skewed Distribution

- Average, median, and mode are **different**
  - The few large (or small) values influence the mean more than the median
  - The highest point is not in the center



Which summary to use?

- Average
  - Best for normal data
  - Preserves totals
- Median
  - Good for skewed data or data with outliers, provided you do not need to preserve or estimate total amounts
- Mode
  - Best for categories (nominal data).
  - The mode is the only summary computable for nominal data!

Which Summary? (continued)

- Average requires quantitative data (numbers)
- Median works with quantitative or ordinal
- Mode works with quantitative, ordinal, or nominal

	Quantitative	Ordinal	Nominal
Average	Yes	-	-
Median	Yes	Yes	-
Mode	Yes	Yes	Yes

Dispersion or Variability

- Also known as dispersion, spread, uncertainty, diversity, risk
- Example data: 2, 2, 2, 2, 2, 2, 2
  - Variability = 0
- Example data: 1, 3, 2, 2, 1, 2, 3
  - How much variability?
  - Look at how far each data value is from average $\bar{X} = 2$:
  - Deviations from average are -1, 1, 0, 0, -1, 0, 1
  - Variability should be between 0 and 1

Dispersion: Examples

- Stock market, daily change, is uncertain
  - Not the same, day after day!
- Risk of a business venture
  - There are potential rewards, but possible losses
- Uncertain payoffs and risk aversion
  - Which would you rather have
    - $1,000,000 for sure
    - $0 or $2,000,000, each outcome equally likely
  - Both have same average! ($1,000,000)
  - Most would prefer the choice with less uncertainty

Dispersion: Standard Deviation *S*

- Measures variability by answering:
  - "Approximately **how far from average** are the data values?" (same measurement units as the data)
  - The square root of the average squared deviation
    - (dividing by *n*-1 instead of *n* for a sample)
- For a sample

$$S=\sqrt{\frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\ldots+(X_n-\bar{X})^2}{n-1}}$$

- For a population

"sigma"

$$\sigma=\sqrt{\frac{(X_1-\mu)^2+(X_2-\mu)^2+\ldots+(X_N-\mu)^2}{N}}$$

Dispersion Example: Spending

- Customers plan to spend ($thousands)

  3.8, 1.4, 0.3, 0.6, 2.8, 5.5, 0.9, 1.1

- Average is 2.05. Sum of squared deviations is

  $(3.8–2.05)^2+(1.4–2.05)^2+…+(1.1–2.05)^2 = 23.34$

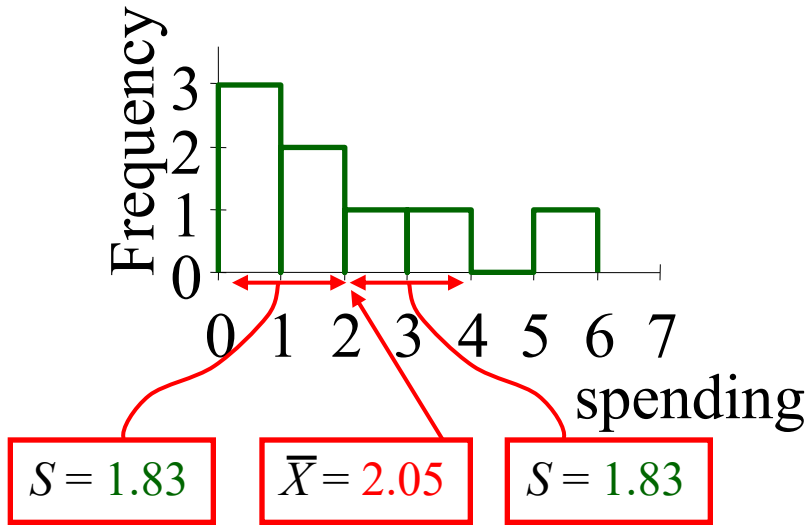
- Divide by 8–1=7 and take square root:

$$\sqrt{\frac{23.34}{7}} = \sqrt{3.334286} = \boxed{1.83} \text{ = Standard deviation}$$

- Customers plan to spend about 1.83 (thousand, i.e., $1,830) more or less than the average, 2.05.
  - Some plan to spend more, others less than average

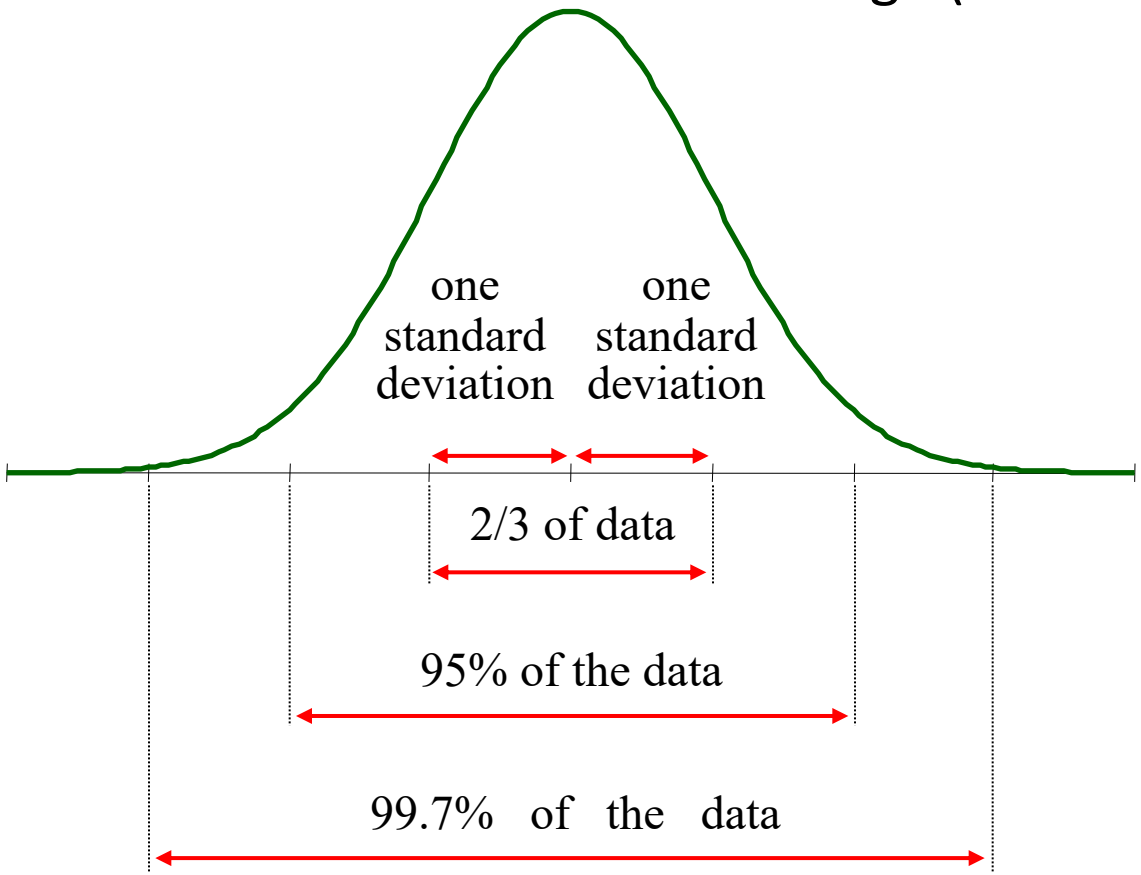
Dispersion Example: Spending

- On the histogram
  - Average is located near the center of the distribution
  - Standard deviation is a *distance* away from the average
  - Standard deviation is the *typical distance* from average



Dispersion in Normal Distribution

- For a **normal distribution only**
  - 2/3 of data within one standard deviation of the average (either above or below)
  - 95% for 2 std. devs.
  - 99.7% for 3



Dispersion in Skewed Distribution

- No simple rule for percentages within one, two, three standard deviations of the average

- Standard deviation retains its interpretation as the standard measure of

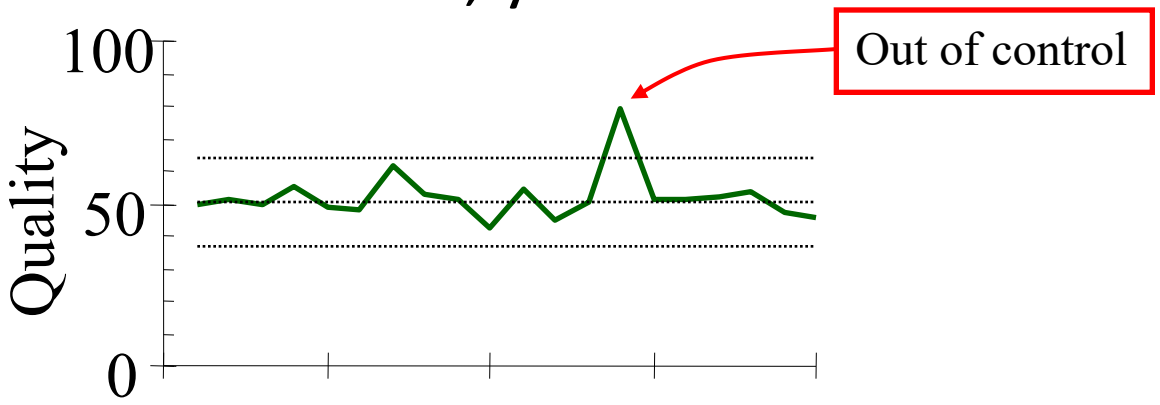
Typically how far the observations are from average

Dispersion Example: Quality Control Charts

- Control limits are often set at

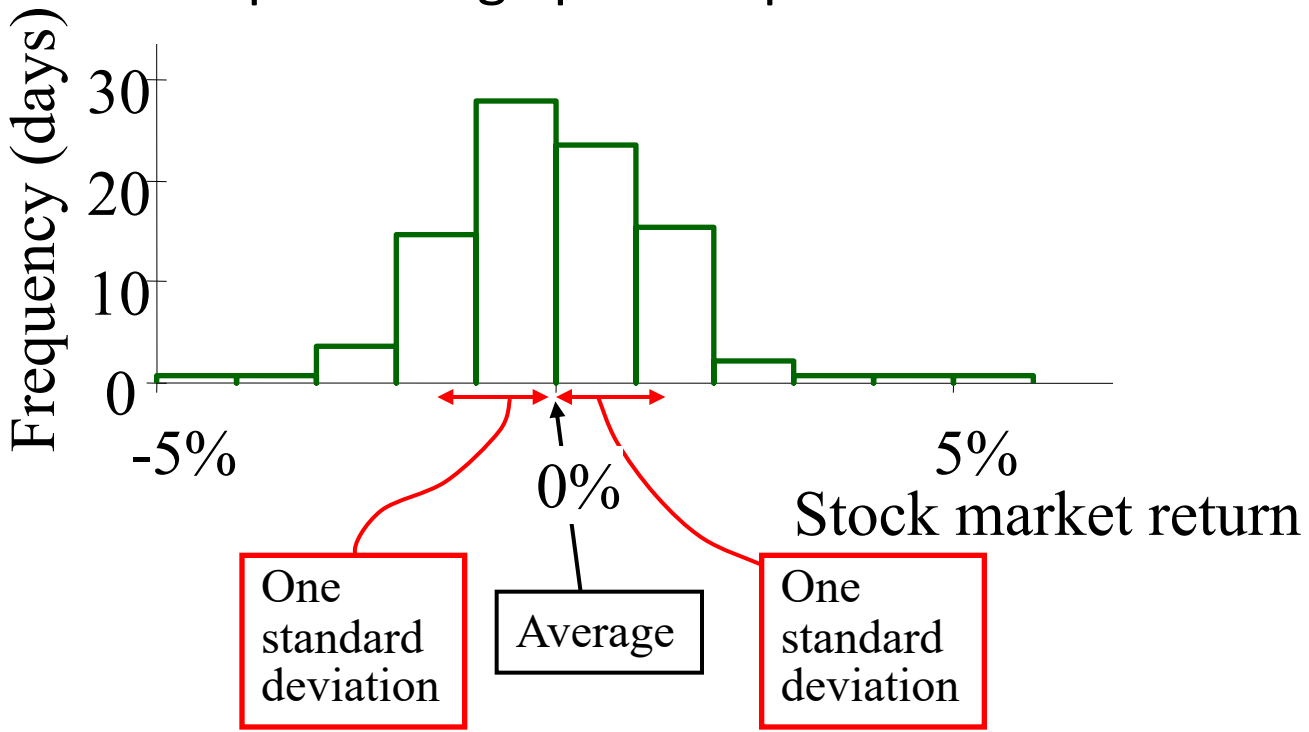
  3 standard deviations from the average

- If the process is normally distributed, then
  - Over the long run, observations will stay within the control limits 99.7% of the time
- If the process goes out of control, you will know



Dispersion in Example: The Stock Market

- Daily stock market returns, S&P500 index, first half of 2001. Standard deviation is 1.43%
  - Average daily percent change: -0.03%
  - Typical day: about 1.5 percentage points up or down

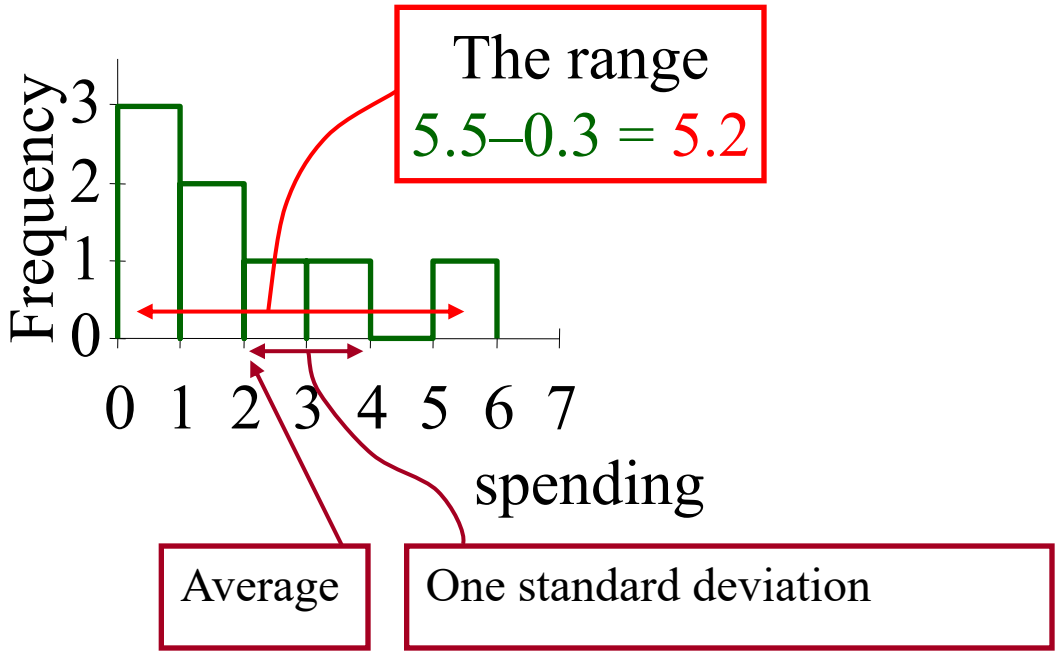


Dispersion: The Range

- The difference: Largest – Smallest
- Good features
  - Easy and fast to compute
  - Describe the data
  - Check the data: Is the range too big to be reasonable?
- Problem
  - Very sensitive to just two data values
    - Compare to standard deviation, which combines all data values

Dispersion Example: Spending

- $Thousands: 3.8, 1.4, 0.3, 0.6, 2.8, 5.5, 0.9, 1.1
- The range is 5.2
  - larger than the standard deviation, 1.83



Dispersion: Coefficient of Variation

- A ***relative*** measure of variability
- The ratio: Standard deviation divided by average
  - For a sample: S/X
  - For a population: σ/μ
- No measurement units. A pure number. Answers:
  - "Typically, in percentage terms, how far are data values from average?"
- Useful for comparing situations of different sizes
  - To see how variability compares *after adjusting for size*

Statistical Inference: Random Sampling

Random Sampling

- The basis for statistical inference about a population based on a sample
  - Example: Build restaurant in neighborhood?
- Population: the collection of items you want to understand

  N items. Example: all people in neighborhood
- Sample: a smaller collection of population units

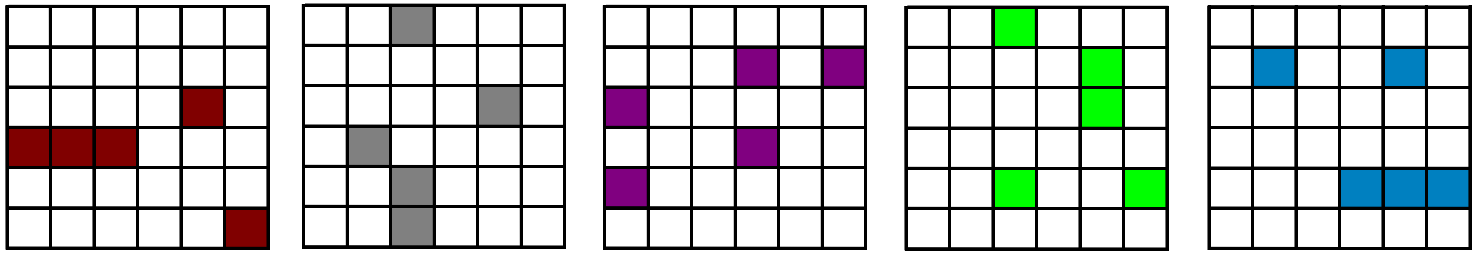
  n items. Example: 100 neighborhood residents who agree to be interviewed
  - Which 100 residents?
  - How to select?

Random Sampling

- Provides a foundation for statistical inference
- A random sample must satisfy:
  1. Each population unit must have an equal chance of being selected
     - This helps assure representation, because all units in the population are equally accessible
  2. Units must be selected independently of one another
     - This guarantees that each item to be selected will bring new, independent information
- Properties
  - Sample is representative of population (on average)
  - Statistical inference will use randomness of the sample

Random Samples of 5 from 36

- On 6 by 6 grid
  - 5 squares shaded in, selected at random, one at a time
    - Without replacement
  - Note that adjacent (touching) squares can be selected
    - To exclude them would make selection non-independent
  - Perhaps you see systematic patterns
    - They are not there by design, but by coincidence
    - But a “checkerboard” pattern would very likely *not* be random



Random Sampling

- Sampling Without Replacement
  - No unit may appear more than once in the sample
    - After a unit is selected, it is removed from the population before another population unit is drawn
- Sampling With Replacement
  - A unit can be represented more than once in the sample
    - After a unit is selected, it is "replaced" back in the population before another unit is drawn
- Census
  - A sample that consists of the *entire population*
    - Often too expensive
    - Even when affordable, may not be cost-efficient

Statistic and Parameter

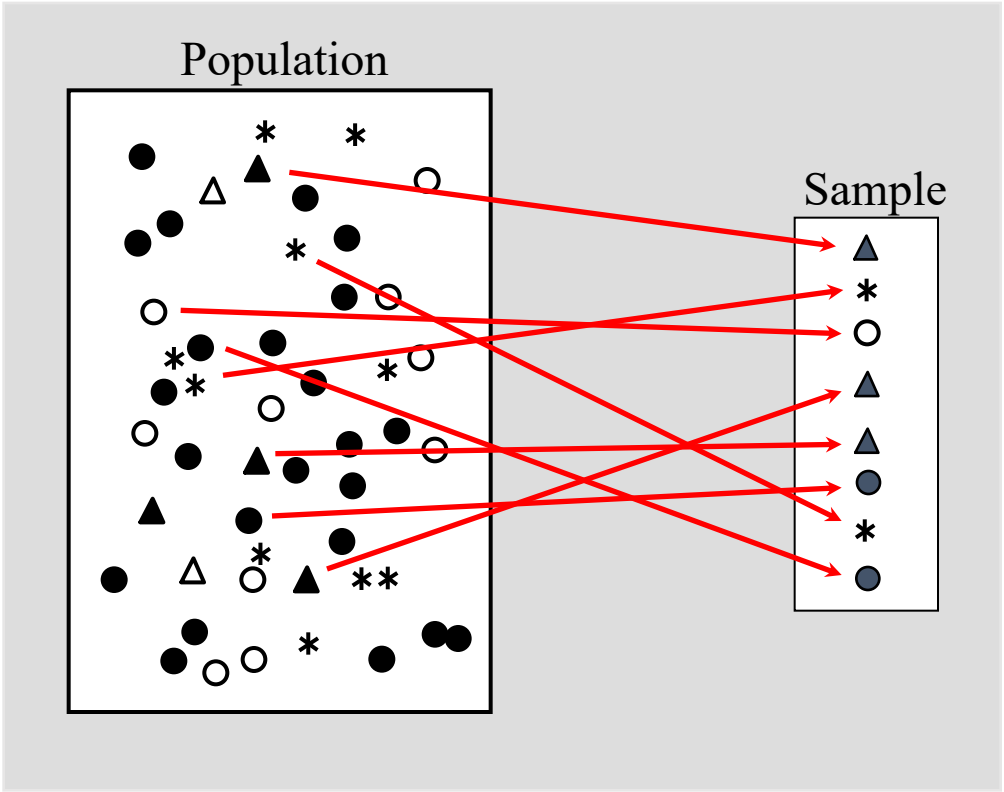
- Sample Statistic
  - Any number computed from sample data
    - A random variable. Known
      - Example: Average weekly food expenditures for 100 sampled residents
        - Random? Yes! Due to randomness of sample selection
- Population Parameter
  - Any number computed for the entire population
    - A fixed number. Unknown
      - Example: mean weekly food expenditures for all 77,386 residents
        - Do we ever know this? NO!
        - But we estimate it (with error)

Estimator and Estimate

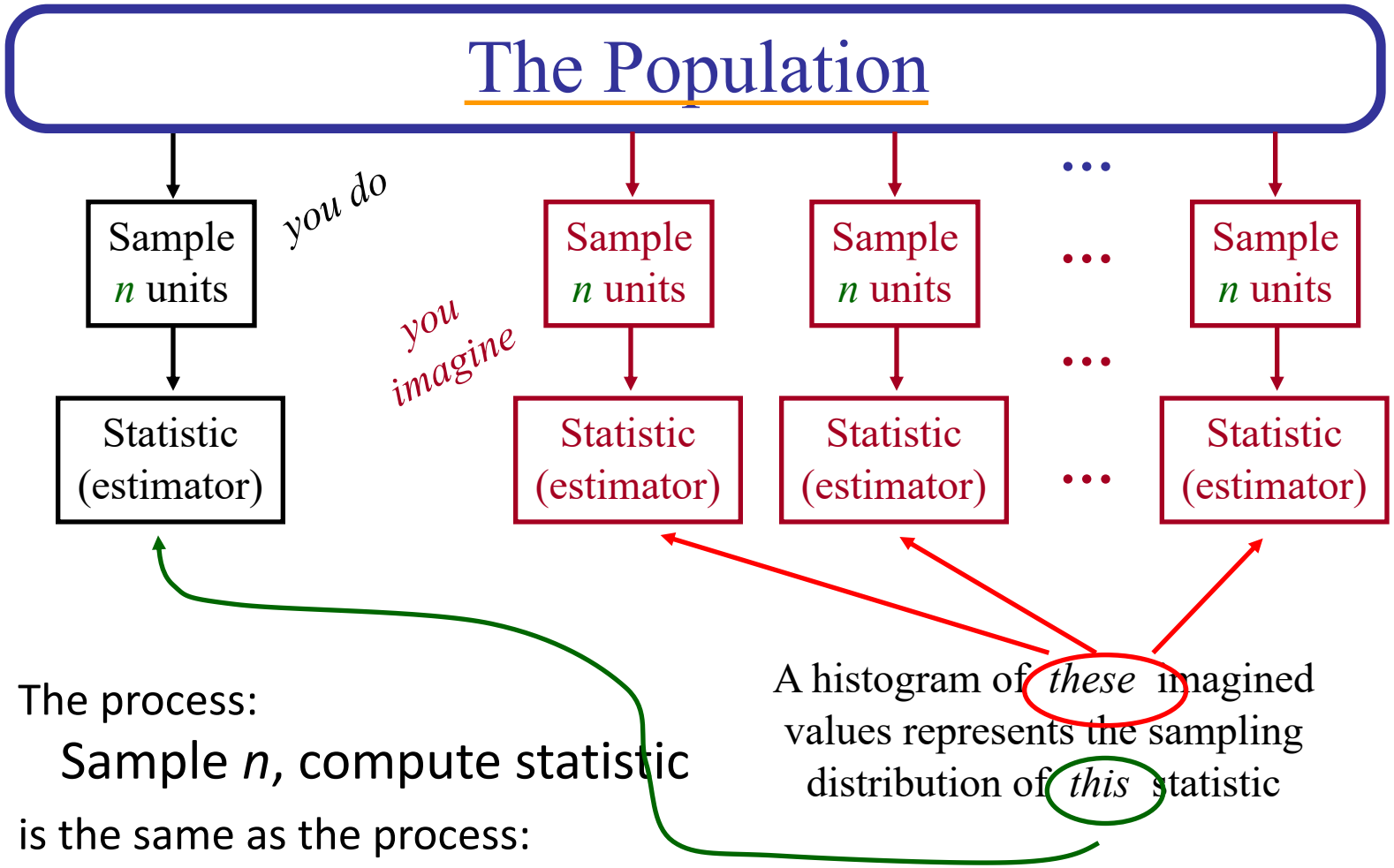
- Estimator
  - A sample statistic used to guess a population parameter
    - Example: Sample average for 100 selected residents is an estimator of the population mean of all 77,386 residents
- Estimate [WRONG! Estimators are usually wrong. Often useful anyway]
  - The actual number computed from the data
    - Example: Rs 33.91 is an estimate of neighborhood weekly food expenditures per person
- Estimation error
  - Estimator minus population parameter. Unknown
    - Example: 33.91 – 35.69 = –1.78

Visualizing the Sampling Process

- Here is a fairly (but not perfectly) representative sample
  - Note that neither open triangle was selected



Visualizing the Sampling Process



- The process:

  Sample *n*, compute statistic

  is the same as the process:

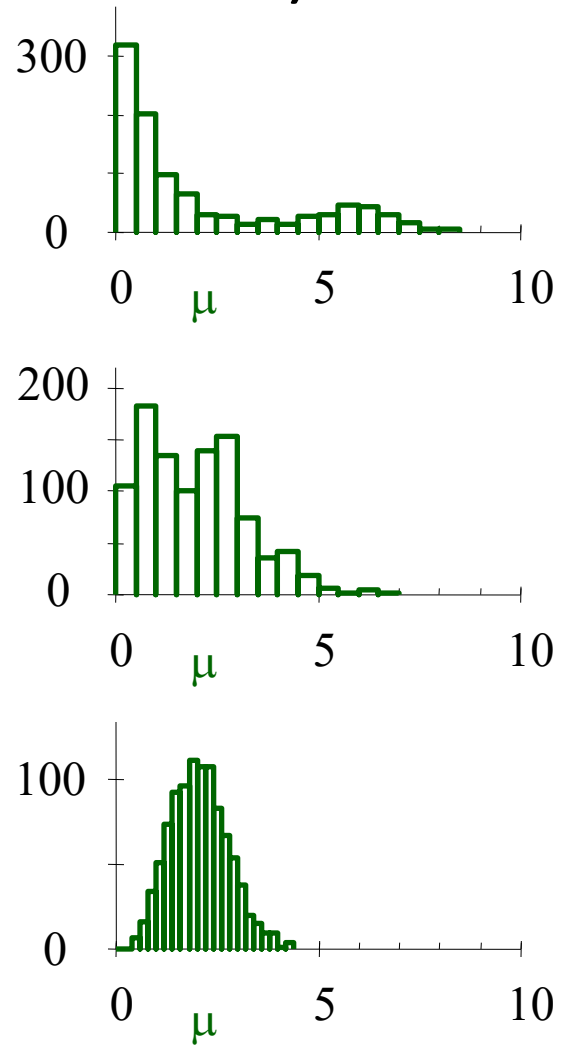
  Choose 1 from sampling distribution

Central Limit Theorem

- Helps you find probabilities for an *average* $\overline{X}$ of *n* independent individuals by giving you
  - The mean $\mu_{\overline{X}} = \mu_X = \mu$
  - The standard deviation $\sigma_{\overline{X}} = \sigma / \sqrt{n}$
  - The right to use the normal probability tables:
    - If *n* is large, then the average $\overline{X}$ is approximately normal, even if individuals are skewed
- Works for totals also by giving you
  - The mean $\mu_{total} = n\mu$
  - The standard deviation $\sigma_{total} = \sigma\sqrt{n}$
  - The right to use the normal probability tables for *total*

Central Limit Theorem (continued)

- Individuals in population
  - Highly non-normal distribution
    - Mean μ, standard deviation σ
- Averages of $n = 3$ individuals
  - Non-normal, but less so
    - Same mean μ
    - Lower std. deviation: $\sigma_{\bar{X}} = \sigma / \sqrt{3}$
- Averages of $n = 10$ individuals
  - Close to normal
    - Same mean μ
    - Lower std. deviation $\sigma_{\bar{X}} = \sigma / \sqrt{10}$



Central Limit Theorem

- What good to us is $\sigma_{\bar{X}} = \sigma / \sqrt{n}$?
  - Why bother choosing a larger sample to estimate μ?
    - Even sampling just *n* = 1 (or a few) we have an *unbiased* estimator that is, on average, equal to μ
      - But the error can be very large!
    - When we sample *n* = 100 or *n* = 1,000, what we gain for our trouble is the LOWER ERROR $\sigma_{\bar{X}} = \sigma / \sqrt{n}$, which says that our estimator, $\bar{X}$, is closer to the unknown population mean μ
  - It tells us how the variability of the sample average $\bar{X}$ is related to the variability of individuals in the population
    - This will be useful soon in defining the *standard error*

Standard Error

- Definition: the Estimated Standard Deviation (of the sampling distribution) of a statistic
- Standard Error of the Average $S_{\bar{X}} = \frac{S}{\sqrt{n}}$
  - Indicates approximately how far the sample average $\bar{X}$ is from the population mean μ
  - Advantage: can be computed using sample data
- Standard Deviation of the Average $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
  - Also indicates approximately how far the sample average $\bar{X}$ is from the population mean μ
  - Problem: cannot be computed without population parameters

Confidence Interval and Hypothesis Testing

Overview

- Confidence Interval
  - Computed from *data*
  - Has a *known* probability of including the *unknown* population parameter being estimated
- Statistical Inference
  - An exact probability statement about the population, based on sample data
- Confidence Level
  - The probability of including the population parameter within the confidence interval
    - **95%** is the usual standard. Also: 99%, 99.9%, 90%

Approximate Confidence Interval

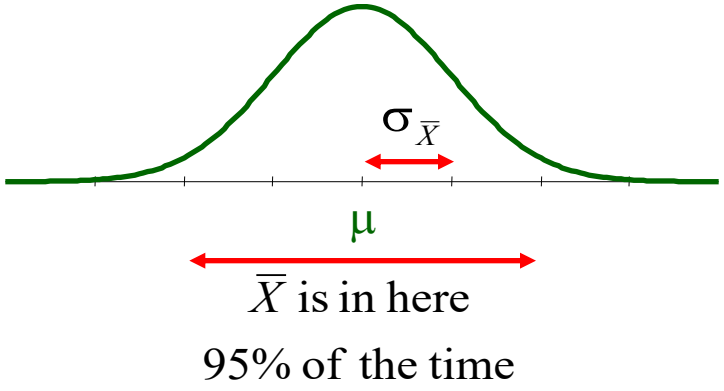
- Estimator ± 2(Standard error of estimator)
  - We are 95% sure the population parameter is in this interval
- For population mean:

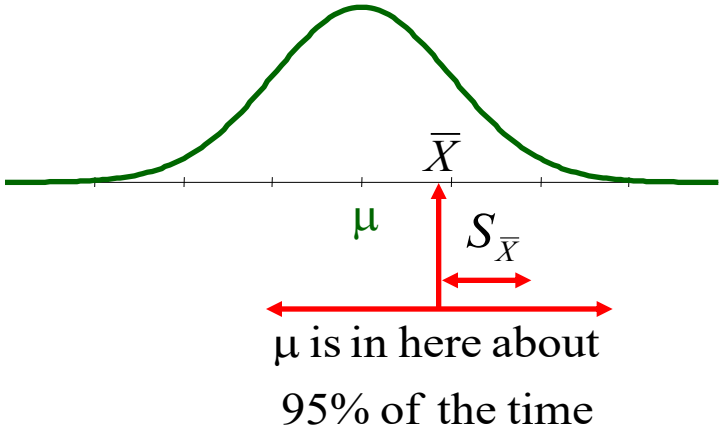
$$\overline{X} \pm 2S_{\overline{X}}$$

- Because an estimator (if normally distributed) has a 95% chance of being within two of its standard deviations from its mean

Why does it work?

- $\bar{X}$ is within $2S_{\bar{X}}$ of its mean μ about 95% of the time
  - Because $S_{\bar{X}}$ is an estimator of $\bar{X}$ (the standard deviation of the sampling distribution of $\sigma_{\bar{X}}$)





- This also says that μ is within $2S_{\bar{X}}$ of $\bar{X}$ about 95% of the time

Confidence Interval for the Mean

- We are 95% sure that the *unknown* population mean μ is between

$$\overline{X} - tS_{\overline{X}} \quad \text{and} \quad \overline{X} + tS_{\overline{X}}$$

- where *t* is from the *t* table
- For smaller sample sizes (*n* = 40 or less)

  t is larger than 1.96 (approximately 2)

  - Because we used the Standard Error ($S_{\overline{X}}$) as an estimator in place of the Population Standard Deviation ($\sigma_{\overline{X}}$) of the sample average

Assumptions

- Assumptions needed for validity of the Confidence Interval

1. Data are a RANDOM SAMPLE from the population of interest
   - (So that the sample can tell you about the population)
2. The sample average $\overline{X}$ is approximately NORMAL
   - *Either* the data are normal (check the histogram)
   - *Or* the central limit theorem applies:
     - Large enough sample size *n*, distribution not too skewed
   - (So that the *t* table is technically appropriate)

Interpretation

- Interpreting a Confidence Interval
  - Population mean μ is fixed and unknown
  - Confidence interval is random and known
  - The probability is 0.95 that μ is between

$$\overline{X} - tS_{\overline{X}} \quad \text{and} \quad \overline{X} + tS_{\overline{X}}$$

  - We are 95% sure that (for example) μ is between

57.86 and 78.74

  - Lifetime average: about 95% of confidence intervals included μ.
    - You may never know which ones!

Hypothesis Testing

- Deciding between two possibilities based on data
  - e.g., “Is it real? Or is it just coincidence?”
- Hypothesis: a statement about the population
  - e.g., More than 30% of customers recognize our product
  - e.g., You will win the election
  - e.g., Strategy *Z* will make you rich in the stock market
- Note: a hypothesis is either ***TRUE*** or ***FALSE***
  - Even with data, you may never know for sure, because of *randomness*

Example: Dishwasher Detergent

- From a box of Cascade: A hypothesis (italics added)
  - "Individual packages of Cascade may weigh slightly more or less than the marked weight due to normal variations incurred with high speed packaging machines, but ***each day's production of Cascade will average slightly above the marked weight***"
- This hypothesis is either ***true*** or ***false***
  - We do not know which
  - The package claims that it is ***true***
  - We could test it, e.g., by weighing a sample of boxes

Example: Pure Randomness

- Coin tossing: Probability 1/2 for Heads or Tails
  - 2 tosses: Probability 1/4 for each of HH, HT, TH, TT
  - 3 tosses: Probability 1/8 for each of
    - HHH, HHT, HTH, HTT, THH, THT, TTH, TTT
    - HHH is no less likely than any other *particular* sequence
  - 10 tosses: probability 1/1,024 of **HHHHHHHHHH**
- If you toss 1,000 times, it is not surprising to find ten Heads in a row *somewhere* in the sequence
- But if you toss only 10 times and find ten Heads
  - You will rightly be suspicious!!!!!
  - Reject the hypothesis that this is an ordinary coin?

Null and Research Hypothesis

Null Hypothesis H_0	Research Hypothesis H_1
The **Default**. Accept unless disproven	Has **burden of proof**. Requires convincing evidence
Often **specific**	Often **general**
Often **randomness**	Often "**your theory**"
A, B are **independent** (no connection)	A, B are **dependent** (related)
Ad has **no effect** on purchase	Ad **works**
$\mu = \mu_0$ Population mean **equals** reference value	$\mu \neq \mu_0$ Population mean **does not equal** reference value
Long-run mean oven temperature, μ, **equals** the desired setting $\mu_0=325^0$	Long-run mean oven temperature, μ, **does not equal** the desired setting $\mu_0=325^0$

Interpretation

- If you ***reject H_0*** and ***accept H_1***
  - H_0 could ***not*** reasonably have produced the data
  - Either
    - H_1 is true, or
    - H_0 is true, but you made a **TYPE I ERROR**
      - Happens 5% of the time when H_0 is true
  - A ***strong conclusion***
  - A ***significant*** result
  - You have earned a "license to explain" the observed difference

Reject H_0 when H_0 is really true

Interpretation (continued)

- If you ***accept H_0***
  - H_0 could reasonably have produced the data
  - Either
    - H_0 is true, or
    - H_1 is true, but you made a **TYPE II ERROR**
      - Difficult to control
      - This error is possible, and is very likely if μ_0 is close to μ
  - A ***weak conclusion***
  - ***Not*** a significant result
  - Little or nothing to explain
  - The observed difference might just be random

Accept H_0 when H_1 is really true

Errors in Hypothesis Testing

The Truth	Your Decision: Accept Null Hypothesis H_0	Your Decision: Accept Research Hypothesis H_1
Null Hypothesis H_0	Yay! Correct Decision	Whoops! Type **I** Error [level 0.05]
Research Hypothesis H_1	Whoops! Type **II** Error [not easily controlled]	Yay! Correct Decision

p-Values

- The smallest test level that is significant
  - Often provided by computer analysis
    - e.g., $p = 0.0297$

Significant, because it is less than 0.05

- Tells the strength of the evidence against H_0
  - Small *p* value says data unlikely to come from H_0
  - Reject H_0 if *p* is small enough
    - Not significant ($p > 0.05$)
    - Significant ($p < 0.05$)
    - Highly significant ($p < 0.01$)
    - Very highly significant ($p < 0.001$)

Says that if H_0 were true, your data would not be surprising

  - What if $p = 0.374$? *Not significant* because $p > 0.05$

Hypothesis Testing for Ordinal data

Nonparametric Methods

- Statistical Procedures for Hypothesis Testing that ***do Not Require a Normal Distribution***
  - Because they are based on ***Counts*** or ***Ranks***
  - A ***Random Sample*** is still required
- The Nonparametric Approach Based on ***Counts***
  - Count the number of times some event occurs
  - Use the binomial distribution to decide whether this count is reasonable or not under the null hypothesis
- The Nonparametric Approach Based on ***Ranks***
  - Replace each data value with its rank (1, 2, 3, …)
  - Use formulas and tables created for testing ranks

Advantages and Disadvantages

- Advantages of Nonparametric Testing
  - No need to assume normality
  - Avoids problems of transformation (e.g., interpretation)
  - Can be used with ordinal data
    - Because ranks can be found
  - Can be much more efficient than parametric methods when distributions ***are not*** normal
- Disadvantage of Nonparametric Testing
  - Less statistically efficient than parametric methods when distributions ***are*** normal
    - Often, this loss of efficiency is slight

Testing the Median

- Testing the Median against a Known Reference Value
  - ***Without*** assuming a normal distribution
  - Note: The number of sample data values below a continuous population's median follows a binomial distribution where $p = 0.5$ and n is the sample size
  - The ***Sign Test***
    1. Find the ***modified sample size m***, the number of data values ***different from*** the reference value θ_0
    2. Find the limits in the table for this modified sample size
    3. Count how many data values fall below the reference value
    4. Significant if the count (step 3) is outside the limits (step 2)

Sign Test: Hypotheses, Assumption

- Sign Test for the Median (for a Continuous Population Distribution)

  $H_0: \theta = \theta_0$ and $H_1: \theta \neq \theta_0$

  - where θ is the unknown population median and θ_0 is the (known) reference value being tested

- Sign Test for the Median (in General)

  H_0: The probability of being above θ_0 ***equals*** the probability of being below θ_0 in the population

  H_1: These probabilities are ***not equal***

  - where θ_0 is the (known) reference value being tested

- Assumption required:
  - The data set is a random sample from the population

Example: Family Income

- Comparing Local to National Family Income
  - Survey median is $70,547, based on $n = 25$ families
  - National median is $27,735
    - This is the reference value θ_0
  - Performing the sign test
    - Modified sample size is $m = 25$, since all sampled families have incomes different from the reference value
    - Limits from the table are 8 and 17 (for testing at the 5% level with $m = 25$)
    - There are 6 families with income below the reference value
    - Since 6 falls outside the limits (from 8 to 17),

Median family income in the community is significantly higher than the national median

Example: Family Income

Income of Sampled families	Is = 27,735	Is > 27, 735	Is < 27,735
$39,465.00	0	1	0
$80,806.00	0	1	0
$267,525.00	0	1	0
$163,819.00	0	1	0
$58,525.00	0	1	0
$25,479.00	0	0	1
$29,341.00	0	1	0
$96,270.00	0	1	0
$85,421.00	0	1	0
$56,240.00	0	1	0
$14,706.00	0	0	1
$54,348.00	0	1	0
$7,081.00	0	0	1
$137,414.00	0	1	0
$16,477.00	0	0	1
$5,921.00	0	0	1
$187,445.00	0	1	0
$83,414.00	0	1	0
$36,346.00	0	1	0
$19,605.00	0	0	1
$156,681.00	0	1	0
$138,933.00	0	1	0
$70,547.00	0	1	0
$81,802.00	0	1	0
$78,464.00	0	1	0
$70,547.00	0	19	6

Testing for Nominal data

Chi-Squared Tests

- Hypothesis Tests for Qualitative Data
  - Categories instead of numbers
  - Based on counts
    - the number of sampled items falling into each category
  - Chi-squared statistic
    - Measures the difference between ***Actual*** counts and ***Expected*** counts (as expected under the null hypothesis H_0)

$$\text{Chi - squared statistic} = \text{Sum of } \frac{(\text{Observed Count} - \text{Expected Count})^2}{\text{Expected Count}} = \sum \frac{(O_i - E_i)^2}{E_i}$$

where the sum extends over all categories or combinations of categories

  - ***Significant*** if the chi-squared statistic is large enough

Summarizing Qualitative Data

- Use Counts and Percentages, for Example:
  - How many (of people sampled) prefer your product?
  - What percentage of sophisticated product users said they would like to purchase an upgrade?
  - What percentage of products produced today are both
    - designed for export
    - ***and*** imperfect

Testing Population Percentages

- Testing if Population Percentages are Equal to Known Reference Values
  - The chi-squared test for equality of percentages
  - Could a table of observed counts have reasonably come from a population with known percentages (the reference values)?
  - ***The data***: A table indicating the ***count*** for each category for a single qualitative variable
  - ***The hypotheses***:
    - H_0: The population percentages are ***equal*** to a set of known, fixed reference values
    - H_1: The population percentages are ***not equal*** to a set of known, fixed reference values

Testing Percentages (continued)

- ***The expected counts***:
  - For each category, multiply the population reference proportion by the sample size, *n*
- ***The assumptions***:
  1. Data set is a random sample from the population of interest
  2. At least 5 counts are expected in each category
- ***The chi-squared statistic***:

$$\text{Sum of } \frac{(\text{Observed Count} - \text{Expected Count})^2}{\text{Expected Count}} = \sum \frac{(O_i - E_i)^2}{E_i}$$

- ***The degrees of freedom***: Number of categories minus 1
- ***The test result***: ***Significant*** if the chi-squared statistic is larger than the critical value from the table

Independence

- Two Qualitative Variables are ***Independent*** if:
  - knowledge about the value (i.e., category) of one variable does not help you predict the other variable
- For Example: Background Sales Information
  - The two variables are
    - where the customer lives, and
    - the customer's favorite product
  - Independence would say that
    - customers tend to have the same pattern (distribution) of favorite products, regardless of where they live, and
    - where customers live is not associated with which product is their favorite

Testing for Association

- The chi-squared test for independence
  - ***The data***: A table indicating the counts for each combination of categories for two qualitative variables
  - ***The hypotheses***:

    H_0: The two variables are ***independent*** of one another

    H_1: The two variables are ***associated***; they are ***not independent***
  - ***The expected table***:

$$\text{Expected count} = \frac{\left(\begin{array}{c}\text{Count for category}\\ \text{for one variable}\end{array}\right)\left(\begin{array}{c}\text{Count for category}\\ \text{for other variable}\end{array}\right)}{n}$$

    - Tells you what the counts *would have been*, on average, if the variables were independent and there were no randomness

Testing Association (continued)

- ***The assumptions***:
  1. Data set is a random sample from the population of interest
  2. At least 5 counts are expected in each combination of categories
- ***The chi-squared statistic***:
  - $$\text{Sum of } \frac{(\text{Observed Count} - \text{Expected Count})^2}{\text{Expected Count}} = \sum \frac{(O_i - E_i)^2}{E_i}$$ where the sum extends over all combinations of categories
- ***The degrees of freedom***:

$$\left(\begin{array}{c}\text{Number of categories}\\ \text{for first variable}\end{array} - 1\right)\left(\begin{array}{c}\text{Number of categories}\\ \text{for second variable}\end{array} - 1\right)$$

- ***The test result***: ***Significant*** if the chi-squared statistic is larger than the critical value from the table

Example: Market Segmentation

- Data: Rowing Machine Purchases
  - Which model rowing machine was purchased?
    - Basic, Designer, or Complete
  - Which type of customer purchased it?
    - Practical or Impulsive

Observed Counts

	Practical	Impulsive	Total
Basic	22	25	47
Designer	13	88	101
Complete	54	19	73
Total	89	132	221

Example (continued)

- Overall Percentages
  - Divide each count by the overall total, 221
    - e.g., 10.0% = 22/221 were practical customers who purchased basic machines
    - Note: impulsive customers bought most designer machines, while practical customers bought most complete machines

Overall Percentages

	Practical	Impulsive	Total
Basic	10.0%	11.3%	21.3%
Designer	5.9%	39.8%	45.7%
Complete	24.4%	8.6%	33.0%
Total	40.3%	59.7%	100.0%

Example (continued)

- Percentages by Model
  - Divide each count by the total for its model type
    - e.g., 22/47 = 46.8% of the 47 basic machines were purchased by practical customers
    - Note: Practical customers make up 40.3% of all customers, but represent 74.0% of complete-machine purchasers

Percentages by Model

	Practical	Impulsive	Total
Basic	46.8%	53.2%	100.0%
Designer	12.9%	87.1%	100.0%
Complete	74.0%	26.0%	100.0%
Total	40.3%	59.7%	100.0%

Example (continued)

- Percentages by Customer Type
  - Divide each count by the total for its customer type
    - e.g., 22/89 = 24.7% of the 89 practical customers purchased basic machines
    - Note: Of all machines, 33.0% are complete; looking only at practical-customer purchases, 60.7% are complete

Percentages by Customer Type

	Practical	Impulsive	Total
Basic	24.7%	18.9%	21.3%
Designer	14.6%	66.7%	45.7%
Complete	60.7%	14.4%	33.0%
Total	100.0%	100.0%	100.0%

Example (continued)

- Expected Counts
  - Multiply row total by column total, then divide by the overall total of 221
    - e.g., if there were no association between type of customer and type of machine, we would have expected to find 89×47/221 = 18.93 basic machines purchased by practical customers

Expected Counts

	Practical	Impulsive	Total
Basic	18.93	28.07	47.00
Designer	40.67	60.33	101.00
Complete	29.40	43.60	73.00
Total	89.00	132.00	221.00

Example (continued)

- Chi-Squared Statistic (do not use the *total* row or column)

  66.8

- Degrees of freedom

  (Rows – 1)(Column – 1) = (3 – 1)(2 – 1) = 2

- Result
  - If there were no association, such a large chi-squared value would be highly unlikely

  The association between customer type and model purchased is very highly significant ($p < 0.001$)