# Introduction

- **Definition -** A set of **data is** a collection of observed values representing one or more characteristics of some objects or units.
- **Definition -** A **population** is a data set representing the entire entity of interest.

# Example – NORC Survey Data

| Respondent | AGE | SEX | HAPPY | TVHOURS |
|---|---|---|---|---|
| 1 | 41 | 1 | 2 | 0 |
| 9 | 31 | 2 | 1 | 0 |
| 22 | 64 | 2 | 3 | 0 |
| 36 | 26 | 1 | 2 | 2 |

This is a survey data from National Opinion Research Center. This is a survey conducted in 1996 for 2904 households from over 70 questions. This is only part of it.

# Data Resource and Format

- **Definition -** A **smple** is a data set consisting of a portion of a population. (Obtained in a way to represent the population)
- **Definition** – A **census** is the collection of the data from everyone on a population.
- Where we can obtain the data:
  - **Primary** data are collected as the part of study.
  - **Secondary** data are obtained from other resources.

# Example - NORC Survey Data

| Respondent | AGE | SEX | HAPPY | TVHOURS |
|---|---|---|---|---|
| 1 | 41 | 1 | 2 | 0 |
| 9 | 31 | 2 | 1 | 0 |
| 22 | 64 | 2 | 3 | 0 |
| 36 | 26 | 1 | 2 | 2 |

This is a survey data from National Opinion Research Center. This is a survey conducted in 1996 for 2904 households from over 70 questions. This is only part of it.

# Questions from NORC Survey Data Data

- Is this data a census data or sample data?
  - Sample data
- What is the population here?
  - All household in United States
- What is the relationship between the population and the sample?
  - Sample is the subset of the population
- Why can we not collect census data in this situation and in most of other situations?
  - It is too expensive and/or time consuming, may not be possible

# Observations and Variables

- **Data format**
  - Observation(s) – a row in the data file
  - Variables(s) – a column in the data file
- Two types of variables – Qualitative (Categorical) variables and Quantitative variables
- **Qualitative (Categorical) variable** - is a variable that is not numerical. It describes data that fits into categories.
- **Quantitative Variable** – is a variable that is measured on a numeric scale for which meaningful arithmetic operations make sense.
- **Remark – Numbers can be qualitative!**

# Variables – Texas House Data

| Obs | Zip | Age | Bed | Bath | Size | Lot | Exter | garage | fp | Price |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 21 | 3 | 2 | 951 | 64904 | Other | 0 | 0 | 30000 |
| 3 | 4 | 7 | 1 | 1 | 676 | 54450 | Other | 2 | 0 | 46500 |
| 5 | 1 | 51 | 3 | 1 | 1186 | 10857 | Other | 1 | 0 | 51500 |
| 7 | 3 | 8 | 3 | 2 | 1368 | . | Frame | 0 | 0 | 56990 |
| 9 | 1 | 51 | 2 | 1 | 1176 | 6259 | Frame | 1 | 1 | 65500 |

Data: 69 families in a midsized city in east Texas. This is only part of it.

# Questions from Texas House Data

- The zip code of a house: qualitative or quantitative?
  - Qualitative
- The square footage of a house: qualitative or quantitative?
  - Quantitative
- Today's snowfall at Houghton: qualitative or quantitative?
  - Quantitative
- Time spent for MA5701: qualitative or quantitative?
  - Quantitative
- Your desire to study MA5701: qualitative or quantitative?
  - Qualitative

# Types of Quantitative Variables

- **Definition -** A **discrete** variable can assume only accountable number of values.
- **Definition –** A **continuous** variable is one that can take any one of an uncountable number of values in an interval.
- **Continuous Variable** – in real world, they may appear discrete but conceptually take any value in an interval. For example, AGE, generally, we calculate it according to your birthdate not in which minute which hour you were born. Other examples include height, blood pressure.

# Types of Qualitative Variables

- **Definition –** The **ordinal scale** distinguishes between measurements Generally, the relative amounts of some characteristic they process.
- **Definition** – The **nominal scale** identifies observed values by name or classification.
  - Generally, for categorical or qualitative variables
  - Weakest scale

# Example - NORC Survey Data

| Respondent | AGE | SEX | HAPPY | TVHOURS |
|---|---|---|---|---|
| 1 | 41 | 1 | 2 | 0 |
| 9 | 31 | 2 | 1 | 0 |
| 22 | 64 | 2 | 3 | 0 |
| 36 | 26 | 1 | 2 | 2 |

**Respondent:**

A. Qualitative
B. Quantitative
C. Ordinal
D. Nominal

# Example - NORC Survey Data

| Respondent | AGE | SEX | HAPPY | TVHOURS |
|---|---|---|---|---|
| 1 | 41 | 1 | 2 | 0 |
| 9 | 31 | 2 | 1 | 0 |
| 22 | 64 | 2 | 3 | 0 |
| 36 | 26 | 1 | 2 | 2 |

**Age:**

A. Qualitative
B. Quantitative
C. Continuous
D. Discrete

# Example - NORC Survey Data

| Respondent | AGE | SEX | HAPPY | TVHOURS |
|---|---|---|---|---|
| 1 | 41 | 1 | 2 | 0 |
| 9 | 31 | 2 | 1 | 0 |
| 22 | 64 | 2 | 3 | 0 |
| 36 | 26 | 1 | 2 | 2 |

**Sex:**

A. Qualitative
B. Quantitative
C. Ordinal
D. Nominal

# Example - NORC Survey Data

| Respondent | AGE | SEX | HAPPY | TVHOURS |
|---|---|---|---|---|
| 1 | 41 | 1 | 2 | 0 |
| 9 | 31 | 2 | 1 | 0 |
| 22 | 64 | 2 | 3 | 0 |
| 36 | 26 | 1 | 2 | 2 |

**HAPPY:**

A. Qualitative
B. Quantitative
C. Ordinal
D. Nominal

# Example - NORC Survey Data

| Respondent | AGE | SEX | HAPPY | TVHOURS |
|---|---|---|---|---|
| 1 | 41 | 1 | 2 | 0 |
| 9 | 31 | 2 | 1 | 0 |
| 22 | 64 | 2 | 3 | 0 |
| 36 | 26 | 1 | 2 | 2 |

**TVHOURS:**

A. Qualitative

B. Quantitative

C. Continuous

D. Discrete

# Variables – Texas House Data

| Obs | Zip | Age | Bed | Bath | Size | Lot | Exter | garage | fp | Price |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 21 | 3 | 2 | 951 | 64904 | Other | 0 | 0 | 30000 |
| 3 | 4 | 7 | 1 | 1 | 676 | 54450 | Other | 2 | 0 | 46500 |
| 5 | 1 | 51 | 3 | 1 | 1186 | 10857 | Other | 1 | 0 | 51500 |
| 7 | 3 | 8 | 3 | 2 | 1368 | . | Frame | 0 | 0 | 56990 |
| 9 | 1 | 51 | 2 | 1 | 1176 | 6259 | Frame | 1 | 1 | 65500 |

**Zip Code:**

A. Qualitative

B. Quantitative

C. Ordinal

D. Nominal

# Variables – Texas House Data

| Obs | Zip | Age | Bed | Bath | Size | Lot | Exter | garage | fp | Price |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 21 | 3 | 2 | 951 | 64904 | Other | 0 | 0 | 30000 |
| 3 | 4 | 7 | 1 | 1 | 676 | 54450 | Other | 2 | 0 | 46500 |
| 5 | 1 | 51 | 3 | 1 | 1186 | 10857 | Other | 1 | 0 | 51500 |
| 7 | 3 | 8 | 3 | 2 | 1368 | . | Frame | 0 | 0 | 56990 |
| 9 | 1 | 51 | 2 | 1 | 1176 | 6259 | Frame | 1 | 1 | 65500 |

**Age:**

A. Qualitative

B. Quantitative

C. Discrete

D. Continuous

# Variables – Texas House Data

| Obs | Zip | Age | Bed | Bath | Size | Lot | Exter | garage | fp | Price |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 21 | 3 | 2 | 951 | 64904 | Other | 0 | 0 | 30000 |
| 3 | 4 | 7 | 1 | 1 | 676 | 54450 | Other | 2 | 0 | 46500 |
| 5 | 1 | 51 | 3 | 1 | 1186 | 10857 | Other | 1 | 0 | 51500 |
| 7 | 3 | 8 | 3 | 2 | 1368 | . | Frame | 0 | 0 | 56990 |
| 9 | 1 | 51 | 2 | 1 | 1176 | 6259 | Frame | 1 | 1 | 65500 |

**Bed or Bath or Garage or Fire Place:**

A. Qualitative

B. Quantitative

C. Discrete

D. Continuous

# Variables – Texas House Data

| Obs | Zip | Age | Bed | Bath | Size | Lot | Exter | garage | fp | Price |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 21 | 3 | 2 | 951 | 64904 | Other | 0 | 0 | 30000 |
| 3 | 4 | 7 | 1 | 1 | 676 | 54450 | Other | 2 | 0 | 46500 |
| 5 | 1 | 51 | 3 | 1 | 1186 | 10857 | Other | 1 | 0 | 51500 |
| 7 | 3 | 8 | 3 | 2 | 1368 | . | Frame | 0 | 0 | 56990 |
| 9 | 1 | 51 | 2 | 1 | 1176 | 6259 | Frame | 1 | 1 | 65500 |

**Size or Lot or Price:**

A. Qualitative

B. Quantitative

C. Discrete

D. Continuous

# Variables – Texas House Data

| Obs | Zip | Age | Bed | Bath | Size | Lot | Exter | garage | fp | Price |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 21 | 3 | 2 | 951 | 64904 | Other | 0 | 0 | 30000 |
| 3 | 4 | 7 | 1 | 1 | 676 | 54450 | Other | 2 | 0 | 46500 |
| 5 | 1 | 51 | 3 | 1 | 1186 | 10857 | Other | 1 | 0 | 51500 |
| 7 | 3 | 8 | 3 | 2 | 1368 | . | Frame | 0 | 0 | 56990 |
| 9 | 1 | 51 | 2 | 1 | 1176 | 6259 | Frame | 1 | 1 | 65500 |

**Exterior Type:**

A. Qualitative

B. Quantitative

C. Ordinal

D. Nominal

# Data – Texas House Data

| Obs | Zip | Age | Bed | Bath | Size | Lot | Exter | garage | fp | Price |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 21 | 3 | 2 | 951 | 64904 | Other | 0 | 0 | 30000 |
| 3 | 4 | 7 | 1 | 1 | 676 | 54450 | Other | 2 | 0 | 46500 |
| 5 | 1 | 51 | 3 | 1 | 1186 | 10857 | Other | 1 | 0 | 51500 |
| 7 | 3 | 8 | 3 | 2 | 1368 | . | Frame | 0 | 0 | 56990 |
| 9 | 1 | 51 | 2 | 1 | 1176 | 6259 | Frame | 1 | 1 | 65500 |

Data: 69 families in a midsized city in east Texas. This is only part of it.

# Distributions

- **Why we need distributions**?
  - Same as other statistics, to use them to summarize data to draw conclusions.
- **Definition –** A **frequency distribution** is a listing of frequencies (counts) of all categories of the observed values of variable.
- **Definition –** A **relative frequency distribution** consists of the relative frequencies, or proportions (percentages), of observations belong to each category.

# Data – Texas House Data

| Obs | Zip | Age | Bed | Bath | Size | Lot | Exter | garage | fp | Price |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 21 | 3 | 2 | 951 | 64904 | Other | 0 | 0 | 30000 |
| 3 | 4 | 7 | 1 | 1 | 676 | 54450 | Other | 2 | 0 | 46500 |
| 5 | 1 | 51 | 3 | 1 | 1186 | 10857 | Other | 1 | 0 | 51500 |
| 7 | 3 | 8 | 3 | 2 | 1368 | . | Frame | 0 | 0 | 56990 |
| 9 | 1 | 51 | 2 | 1 | 1176 | 6259 | Frame | 1 | 1 | 65500 |

Data: 69 families in a midsized city in east Texas. This is only part of it.

# Frequency Table – Discrete Variable

| bed | | | | |
|---|---|---|---|---|
| bed | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 1 | 1 | 1.45 | 1 | 1.45 |
| 2 | 3 | 4.35 | 4 | 5.80 |
| 3 | 46 | 66.67 | 50 | 72.46 |
| 4 | 16 | 23.19 | 66 | 95.65 |
| 5 | 3 | 4.35 | 69 | 100 |

# Frequency Table – Continuous Variables

| price | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| [ 0, 50k) | 4 | 5.80 | 4 | 5.80 |
| [ 50k, 100k) | 22 | 31.88 | 26 | 37.68 |
| [100k, 150k) | 23 | 33.33 | 49 | 71.01 |
| [150k, 200k) | 10 | 14.49 | 59 | 85.51 |
| [200k, 250k) | 2 | 2.90 | 61 | 88.41 |
| [250k, 300k) | 1 | 1.45 | 62 | 89.86 |
| [300k, 350k) | 4 | 5.80 | 66 | 95.65 |
| [350k, 400k) | 3 | 4.35 | 69 | 100.00 |

# Statistics in Chapter 1

- Sample size for each variable
- For a single qualitative variable,
  - Frequency and Relative frequency
- For a single quantitative variable,
  - Sample mean/variance/standard deviation, median, quantiles
- For two qualitative variables
  - Two-dimensional contingency table with frequency
- For a qualitative variable and a quantitative variable
  - Sample mean/variance for each level of qualitative variable

# Importance of Data Displays

- It is a part of **Descriptive Statistics** – organizing and summarizing data through graphics.
- Clever plots of data provide a powerful way for looking at data – sometimes provide more than enough information to answer questions.
- Formal statistical analysis depends on underlying models and assumptions – plots provide a quick, easy, and effective way to check and verify those models and assumptions.
- Generally used Statistical Software can produce a correctly constructed chart and graph with some adjustments

# Plots in Chapter 1

- For a single qualitative variable
  - Barplot
- For a single quantitative variable
  - Boxplot – for small to moderate data
  - Histogram - for moderate and large data
- For a qualitative variable and a quantitative variable
  - Boxplots – boxplot of the quantitative variable for each level of the qualitative variable
- For two quantitative variables
  - Scatter plot

# Introduction of R

- R web site – download and install R
  - https://www.r-project.org/
- Open R
  - Create a data or input a data from a data file to R
  - Perform analysis
- Put R scripts to a file so they can be used later
  - Copy and paste to R

# Basic Data Types in R

- Numeric
  - Such as 1, 10, -0.24, 3.5
- Integer
  - 1L, 55L, -3L
- Complex – not used in our course
- Character (String)
  - “10”, “A”, “Car”
- Logical (Boolean)
  - TRUE or FALSE

# Basic Data Types in R

- Quantitative variables
  - Numeric
  - Integer
- Qualitative variables
  - Factor (a data structure and can be ordered or unordered)
  - Numeric – need to change to factor
  - Character – is considered as factor automatically

# Basic Data Structure in R - Vector

- Vector is an **ordered** collection of **same types** of data with a given length.
- Numeric vector: x <- c(1, 3, 7, 8)
- Character vector: y <- c("A", "B", "MY")
- Here, x <- c(1, 3, 7, 8)
  - A vector (data) named x is created. Equivalently, you can also use x = c(1, 3, 7, 8),
  - c() is a R function that combines its arguments.
  - This vector (x) has 4 elements: 1 (first element), 3 (second element), 7 (third element), 8 (fourth element).

# Names in R

- Can be letters, numbers, ., and _.
  - For example, A1, b2, c.s, c_s
- Cannot start with a number
  - Better start with a letter
- Lower cases and upper cases are different
  - For example, A1 and a2 refer to different data/functions/etc.

# Subscripting with Vector

- x <- 2 * (1:5) or x <- c(2, 4, 6, 8, 10)
- Positive integers
  - x[2] returns 4 (a scalar, which is a vector with length of 1)
  - x[c(1, 3, 5)] returns a vector with three elements: 2, 6, 10
  - x[c(5, 1, 3)] returns a vector with two elements: 10, 2, 6
- Negative integers
  - x[-2] is equivalent to x[c(1, 3, 4, 5)]
  - x[c(-2, -4)] is equivalent to x[c(1, 3, 5)]
  - How about x[c(-4, -2)]?

# Subscripting with Vector

- Cannot combine positive and negative integers
- More on subscripting: x <- c(2, 4, 6, 8, 10)
  - Results of x[c(1, 1)]?
  - Results of x[c(1, 2, 1)]?
- Logical
  - x[c(TRUE, FALSE, FALSE, TRUE, FALSE)] is equivalent to x[c(1, 4)]
  - How about x[TRUE]

# Element-wise Operation

- Arithmetic operation: +, -, *, and /
- Element-wise: it performs each operation on each element of a vector independently of the other elements.
- Two vectors with same length:
  - c(1, 3, 5) + c(2, 4, 6) is c(1 + 2, 3 + 4, 5 +6)
- One vector and one scalar:
  - c(1, 3, 5) + 10 is equivalent to c(1, 3, 5) + c(10, 10, 10)
- Two vectors with different lengths (not covered here)

# Data Structure – Matrix and Data Frame

- Matrix and data frame are a two-dimensional of data in row and columns.
- All rows have the same length
- All columns have the same length
- Matrix
  - All columns must be the same type,
  - If one column is numeric, then all the other columns must be numeric
- Data frame
  - Columns can have different types of data
  - One column is numeric, the other column can be character
- Vector is used to store data with one variable
- Data frame is used to store data with more than one variable

# Subscripting with Matrix/Data Frame

- x <- 2 * (1:5) or x <- c(2, 4, 6, 8, 10)
- Positive integers
  - x[2] returns 4 (a scalar, which is a vector with length of 1)
  - x[c(1, 3, 5)] returns a vector with three elements: 2, 6, 10
  - x[c(5, 1, 3)] returns a vector with two elements: 10, 2, 6
- Negative integers
  - x[-2] is equivalent to x[c(1, 3, 4, 5)]
  - x[c(-2, -4)] is equivalent to x[c(1, 3, 5)]
  - How about x[c(-4, -2)]?

# Input Data from a Data File

- Set up the working directory – the file folder contains your data files
  - Put all data files to a single folder
  - Use the full path and /
  - setwd("G:/My Drive/Zkui/Teaching/DataSets/MA5701")
- Use R function read.table and/or read.csv
  - norc <- read.csv(file = "norc.csv", stringsAsFactors = FALSE)

# Basic Functions to Look at a Data Frame

- Print first or last few rows of a data frame
  - head(norc)
  - tail(norc)
- Name of columns
  - names(norc)
- Structure of a data frame
  - str(norc)

# Subscripting with Matrix/Data Frame

- Subscripting is done with [ for rows, for columns]
- Each part can be positive/negative integers and logical
  - Blank means all rows/all columns
  - First row and all columns: norc[1, ]
  - First and third row and all columns: nocr[c(1, 3), ]
  - Second column and all rows: norc[ , 2]
  - Second and third columns and all rows: norc[ , c(2, 3)]
  - First & third rows and first & second columns: norc[c(1, 3), c(1, 2)]
- Column names for one or more columns
  - Second column: norc$age

# Frequency Table – Discrete Variable

| bed | | | | |
|---|---|---|---|---|
| bed | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 1 | 1 | 1.45 | 1 | 1.45 |
| 2 | 3 | 4.35 | 4 | 5.80 |
| 3 | 46 | 66.67 | 50 | 72.46 |
| 4 | 16 | 23.19 | 66 | 95.65 |
| 5 | 3 | 4.35 | 69 | 100 |

# Construct Frequency Table with R

- Frequency table: table() function
  - table(texas$bed)
- Relative frequency table: prop.table()
  - Input is the output from table()
  - prop.table( table(texas$bed) )
- Use appropriate decimal digits: round()
  - round(b, digits = 4)

# Frequency Table – Continuous Variables

| price | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| [ 0, 50k) | 4 | 5.80 | 4 | 5.80 |
| [ 50k, 100k) | 22 | 31.88 | 26 | 37.68 |
| [100k, 150k) | 23 | 33.33 | 49 | 71.01 |
| [150k, 200k) | 10 | 14.49 | 59 | 85.51 |
| [200k, 250k) | 2 | 2.90 | 61 | 88.41 |
| [250k, 300k) | 1 | 1.45 | 62 | 89.86 |
| [300k, 350k) | 4 | 5.80 | 66 | 95.65 |
| [350k, 400k) | 3 | 4.35 | 69 | 100.00 |

# Construct Frequency Table with R

- For continuous variables: cut() and table()
- Function cut():
  - Divides the range of data into intervals and codes the values according to which interval they fall.
  - a <- cut(x = texas$price,

    breaks = c(0, 50, 100, 150, 200, 250, 300, 350, 400) * 1000,

    include.lowest = TRUE, right = FALSE)
  - Then use table(a) etc.

# Bar plot (bar chart)

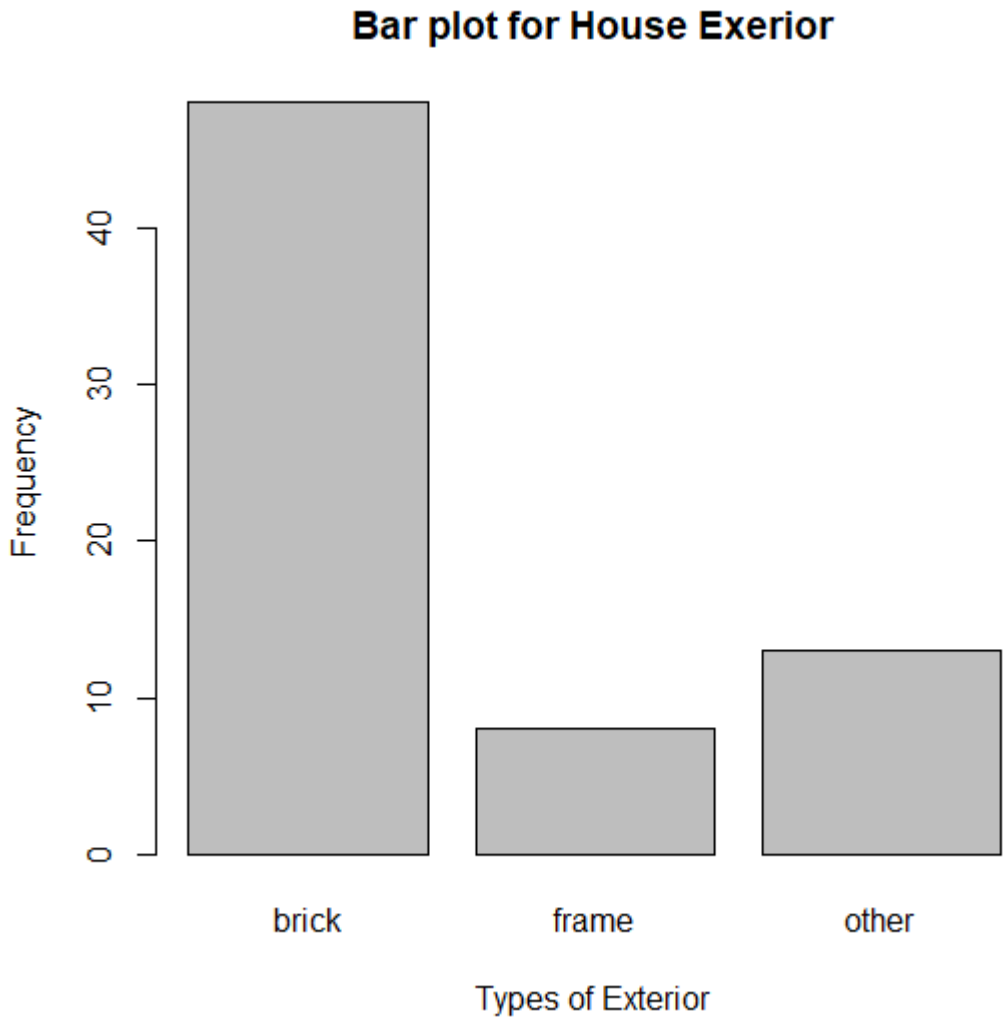
- A **bar plot** is a plot that uses the height of rectangles (bars) to represent the frequency of each value.
- Look for differences in the heights of the bars.
- R function: barplot()
  - Use the output of table() as the input
  - For example, barplot( table(texas$exter) )

# Bar plot (bar chart)

- A **bar plot** is a plot that uses the height of rectangles (bars) to represent the frequency of each value.
- Look for differences in the heights of the bars.
- R function: barplot()
  - Use the output of table() as the input
  - For example, barplot(table(texas$exter))

# Bar plot (bar chart)

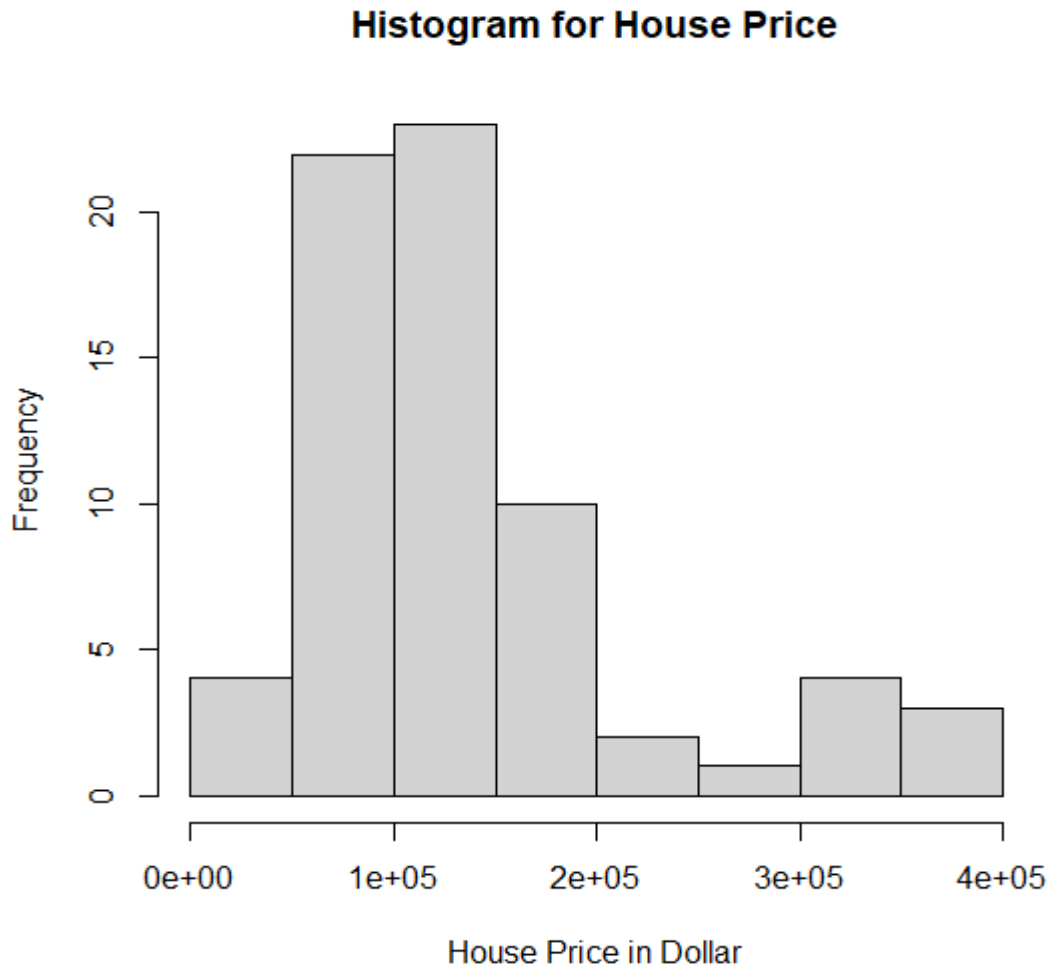
- **Conclusions**
  - Most of houses have the brick exterior
  - About the same number of houses have the frame or other exteriors.



# Histogram

- A **histogram** is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same and the rectangles touch each other.
- We will rely on R function hist() to construct the histogram.

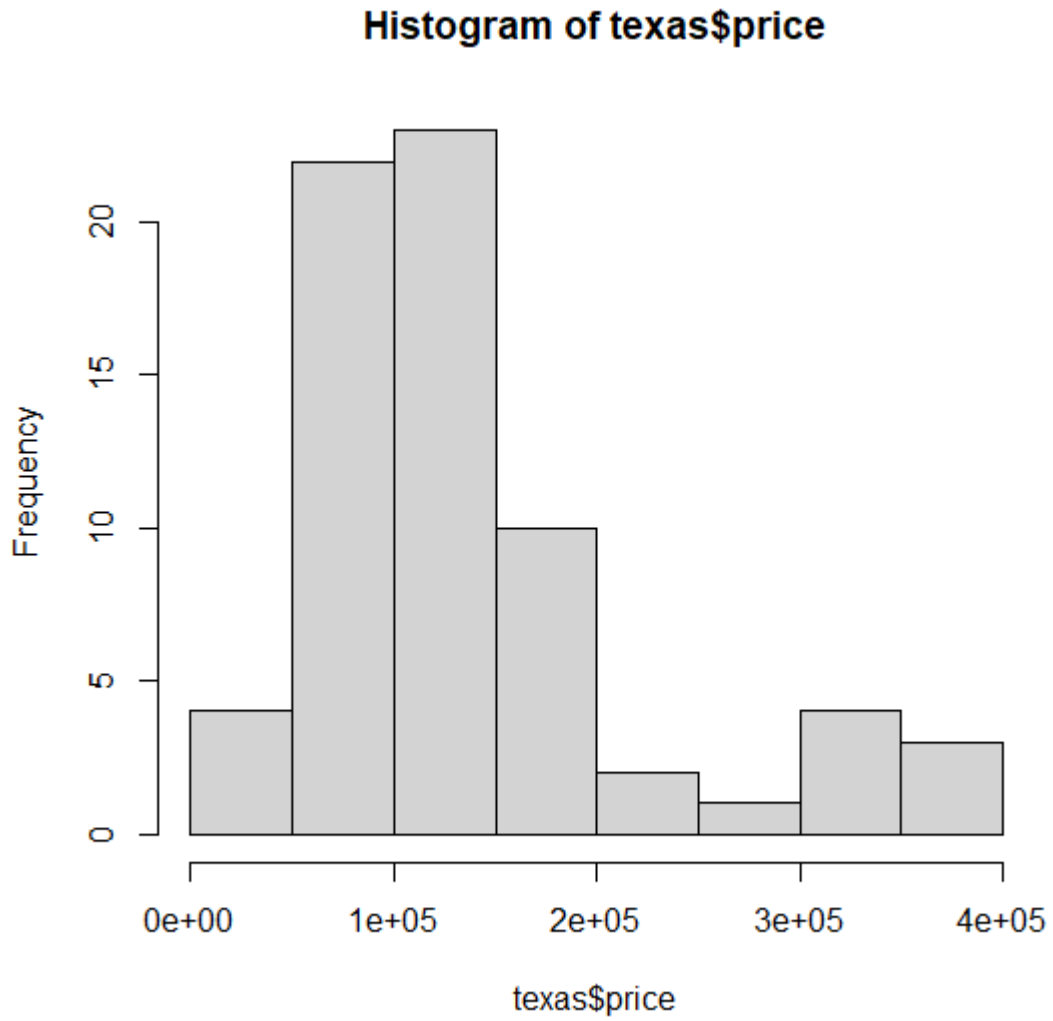
# Histogram from Texas House Data



# Histogram

```
hist(texas$price,
   breaks = c(0, 50, 100, 150, 200, 250, 300, 350, 400) *
        1000,
   right = FALSE,
   main = "Histogram for House Price",
   xlab = "House Price in Dollar",
   ylab = "Frequency")
```

# Histogram from Texas House Data



# Histogram - Interpretation

- We describe the data by the “center” and the “spread”.
  - What is a typical value (“**center**”) of the data?
  - What is the variability (“**spread**”) of the data?
- Do the data follow some **patterns**?
  - Symmetric
  - Skewed-left (left-skewed, negatively skewed, or left-tailed)
    - More values on the right side and the tail on the left side is longer
  - Skewed-right (right-skewed, positively skewed, or right-tailed)
    - More values on the left side and the tail on the right side is longer
- Are there **multiple peaks** – multiple peaks suggest a mixture of populations

# Shape of the Data

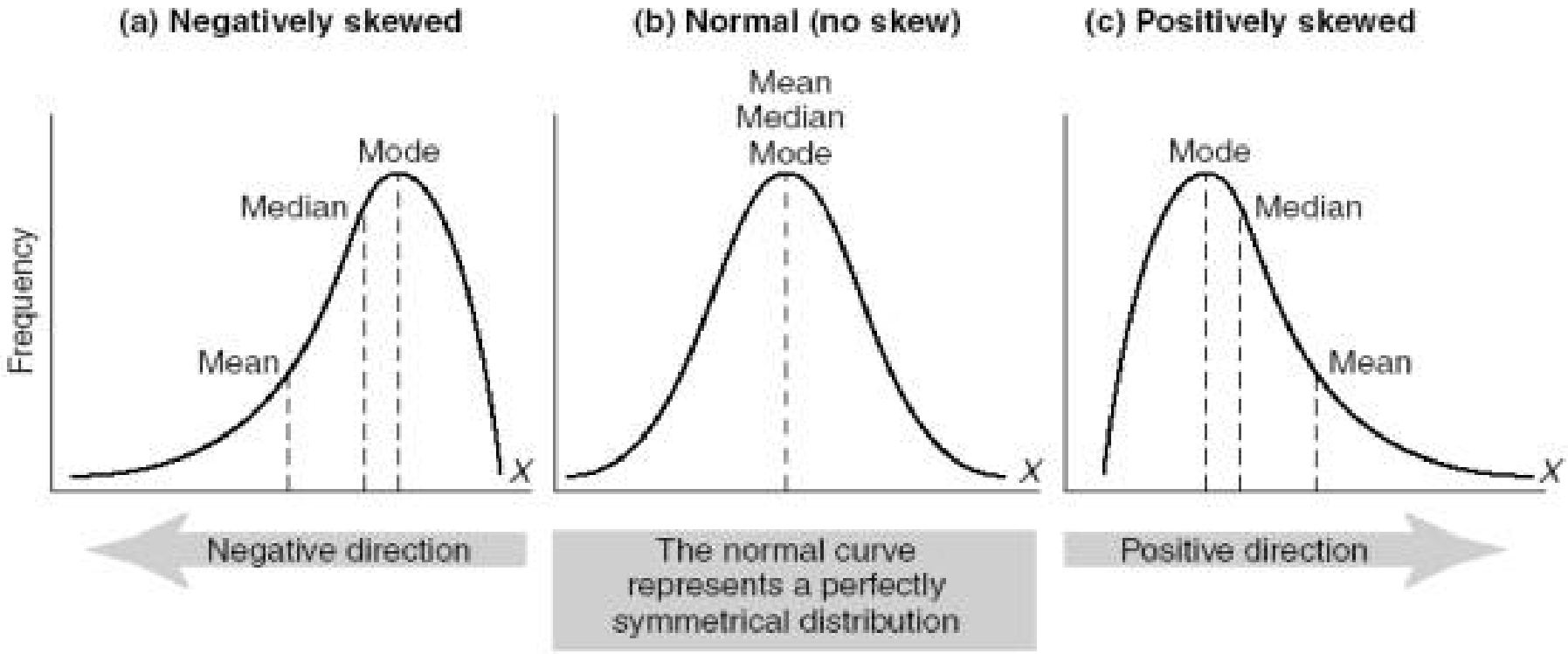
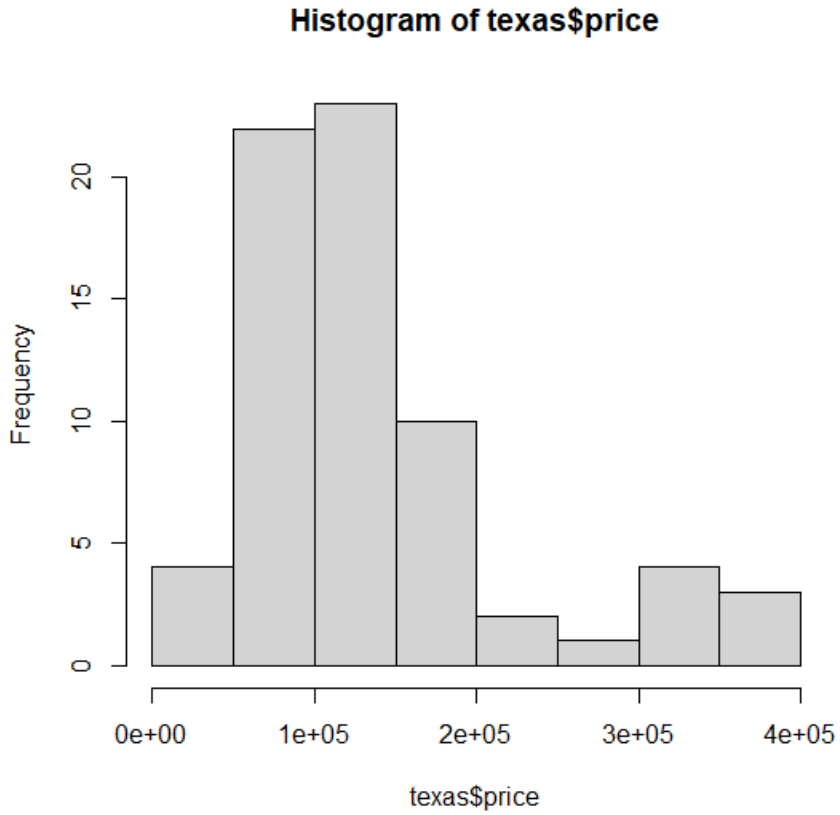



FIGURE 15.6 Examples of normal and skewed distributions

# Histogram – House Price

- **Center:** Most of houses have the price between 50k and 250k
- **Spread:** The house price ranges from 30k to around 400k.
- **Patterns:** The data is not symmetric. The data is right-skewed.
  - This meaning that most houses have a lower price (less than 250k) but a few houses are much more expensive (greater than 350k).
- **Multiple peaks:** There is one peak around 200k.



# Descriptive Statistics - Mean

- **Definition –** The **mean (sample mean)** is the sum of all the observed values divided by the number of values.
- Let $y_1, \cdots, y_n$ denote a sample of interest. The mean (sample mean), denoted by $\bar{y}$, is given by

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n).$$

- Mean is a **Measure of Location or Center Tendency**.

# Descriptive Statistics – Sample Variance

- **Definition –** The **sample variance**, denoted by $s^2$ is defined by

$$s^2 = \frac{1}{n-1}[(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2]$$

- It looks like an "average" of the squared deviations from the sample mean.
- Note that we use $n - 1$ instead of $n$.
- Sample variance is a measure of **dispersion** which is the extent to which a distribution is stretched or squeezed.

# Descriptive Statistics – Sample Variance

- Another formula for sample variance, $s^2$, is

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

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right).$$

- **Definition –** The **standard deviation** of a set of observed values is defined to be the positive square root of the variance. In other words, the sample standard deviation is: $s = \sqrt{s^2}$.

# Mean and Standard Deviation

- **Usefulness of the Mean and Standard Deviation**
  - Interval (mean +- 1*SD) contains approximately 68% of observations
  - Interval (mean +- 2*SD) contains approximately 95% of observations
  - Interval (mean +- 3*SD) contains virtually all of the observations

# Sample Mean/Variance – A Toy Example

- Find sample mean/variance for data: {1, 5, 4, 9, 6}.

$$\sum y_i = 1 + 5 + 4 + 9 + 6 = 25$$

$$\sum y_i^2 = 1^2 + 5^2 + 4^2 + 9^2 + 6^2 = 159$$

$$\sum (y_i - \bar{y})^2 = (-4)^2 + 0^2 + (-1)^2 + 4^2 + 1^2 = 34$$

- Sample mean is: $\frac{1}{n}\sum y_i = \frac{25}{5} = 5$
- Sample variance is: $\frac{1}{n-1}(\sum y_i^2 - n\bar{y}^2) = \frac{159-5*5*5}{4} = 8.5$
- Sample variance is: $\frac{1}{n-1}\sum (y_i - \bar{y})^2 = \frac{34}{4} = 8.5$

# Sample Mean/Variance – Packaged Weights

- **Example** – King (1992) discusses the net weights of a nominally 16-oz packaged product. An inspector collected of 20 packages and measured their net contents.

| | | | | |
|---|---|---|---|---|
| 16.4 | 16.4 | 16.5 | 16.5 | 16.6 |
| 16.7 | 16.2 | 16.4 | 16.4 | 16.5 |
| 16.6 | 16.6 | 16.8 | 16.3 | 16.4 |
| 16.5 | 16.5 | 16.6 | 16.7 | 16.8 |

# Sample Mean/Variance – Packaged Weights

- $\sum_{i=1}^{20} y_i = 16.4 + 16.4 + \cdots + 16.8 = 330.4$
- So sample mean is: $\bar{y} = 330.4/20 = 16.52$
- $\sum_{i=1}^{20} y_i^2 = 16.4^2 + 16.4^2 + \cdots + 16.8^2 = 5458.68$
- So sample variance is
- $s^2 = (5458.68 - 20 * 16.52 * 16.52)/19 = 0.02484$
- $\sum_{i=1}^{20} (y_i - \bar{y})^2 = (16.4 - 16.52)^2 + \cdots + (16.8 - 16.52)^2 = 0.472$
- So sample variance is $s^2 = \frac{0.472}{19} = 0.02484$

# Descriptive Statistics – Percentile (Quantile)

- **Definition –** The **median** of a set of observed values is defined to be the middle value when the measurement are arranged from lowest to the highest.
- **Definition –** The ***p*th percentile** (**quantile**) is defined to be that value for which at most (*p*)% of the measurement are less and at most (100-*p*)% of the measurement are greater.

# Descriptive Statistics - Percentile

- **Quartiles**, 25%, 50%, 75% percentile
  - 25% percentile – lower quartile, first quartile ($Q_1$)
  - 50% percentile – median, second quartile
  - 75% percentile – upper quartile, third quartile ($Q_3$)
- Question: what are 100% percentile and 0% percentile?
  - 100% percentile – maximum or largest
  - 0% percentile – minimum or smallest
- **Definition** – The **interquartile range** is the length of the interval between the $25^{th}$ and $75^{th}$ percentiles.

# Descriptive Statistics - Median

- **Definition –** The **median** of a set of observed values is defined to be the middle value when the measurement are arranged from lowest to the highest.
- Let $y_1, \cdots, y_n$ denote the data and $\tilde{y}$ denote the median.
- Arrange the data in ascending order: $y_{(1)} \leq \cdots \leq y_{(n)}$
- Median $\tilde{y} = y_{(\frac{n+1}{2})}$ if $n$ is odd; $\tilde{y} = \frac{y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)}}{2}$ if $n$ is even.

# A Toy Example

- Suppose we have a data ($n$ = 6): 3, 1, 5, 20, 3, 12, then

  $y_1 = 3, y_2 = 1, y_3 = 5, y_4 = 20, y_5 = 3, y_6 = 12$

- We arrange the data in ascending order: 1, 3, 3, 5, 12, 20, then

  $y_{(1)} = 1, y_{(2)} = 3, y_{(3)} = 3, y_{(4)} = 5, y_{(5)} = 12, y_{(6)} = 20$

- So $y_{(1)}$ is the smallest while $y_{(n)}$ is the largest.

- The median is $\tilde{y} = \frac{y_{(3)} + y_{(4)}}{2} = \frac{3+5}{2} = 4$

# Example – Wall Thickness of Aircraft Parts

- **Example** – Eck Industries, Inc. Manufacturers cast aluminum cylinder heads that used for liquid-cooled aircraft engines. The thicknesses (in inches) for 18 cylinder heads are given below:

| | | | | | |
|---|---|---|---|---|---|
| 0.223 | 0.193 | 0.218 | 0.201 | 0.231 | 0.204 |
| 0.228 | 0.223 | 0.215 | 0.223 | 0.237 | 0.226 |
| 0.214 | 0.213 | 0.233 | 0.224 | 0.217 | 0.210 |

- Find the median of this data.

# Example – Wall Thickness of Aircraft Parts

- First, sort the data to the ascending order:

| | | | | | |
|---|---|---|---|---|---|
| 0.193 | 0.201 | 0.204 | 0.210 | 0.213 | 0.214 |
| 0.215 | 0.217 | 0.218 | 0.223 | 0.223 | 0.223 |
| 0.224 | 0.226 | 0.228 | 0.231 | 0.233 | 0.237 |

- Since $n$ = 18, the median is

$$\tilde{y} = \frac{y_{(9)} + y_{(10)}}{2} = \frac{0.218 + 0.223}{2} = 0.2205$$

# Descriptive Statistics - Quartiles

- We will rely on R to calculate quartiles.

# Descriptive Statistics with R

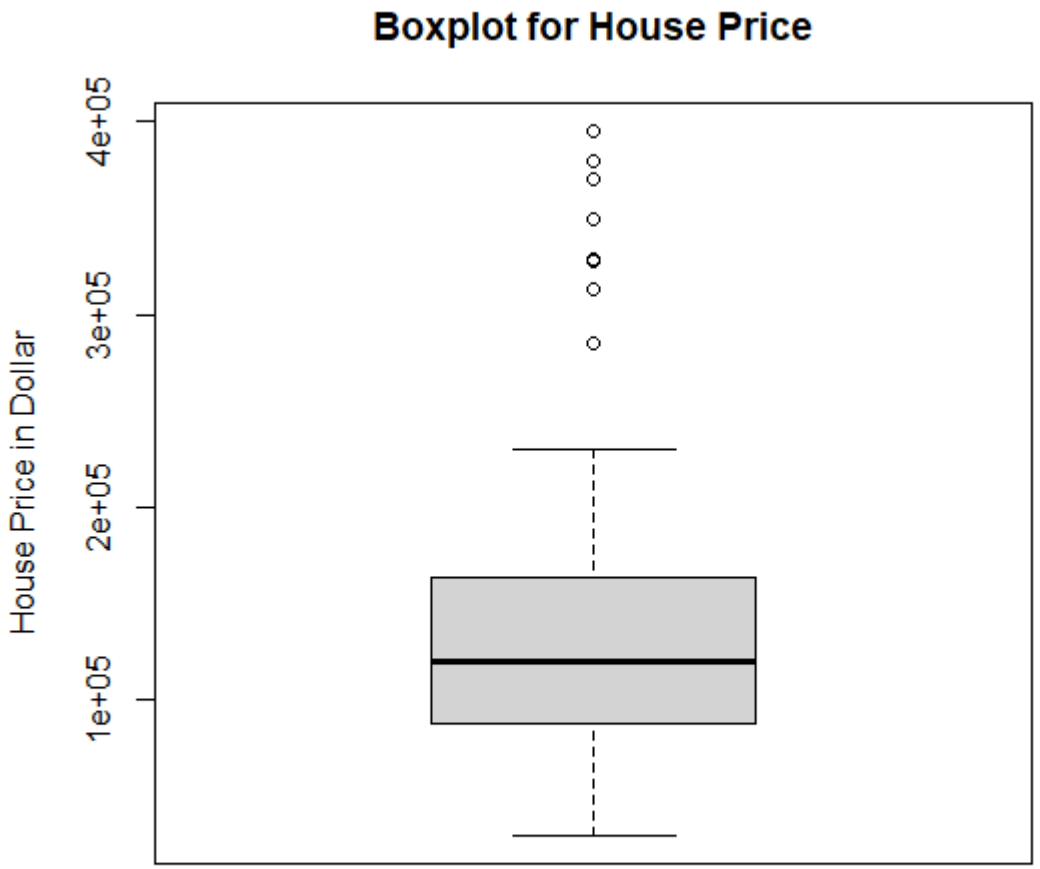
- Let use the data for the age from NORC
- Smallest and largest: min() and max()
- Sample mean: mean()
- Sample variance/standard deviation: var() and sd()
- Median: median()
- Quantile: quantile()

# Boxplots

The boxplot provides the analysis:

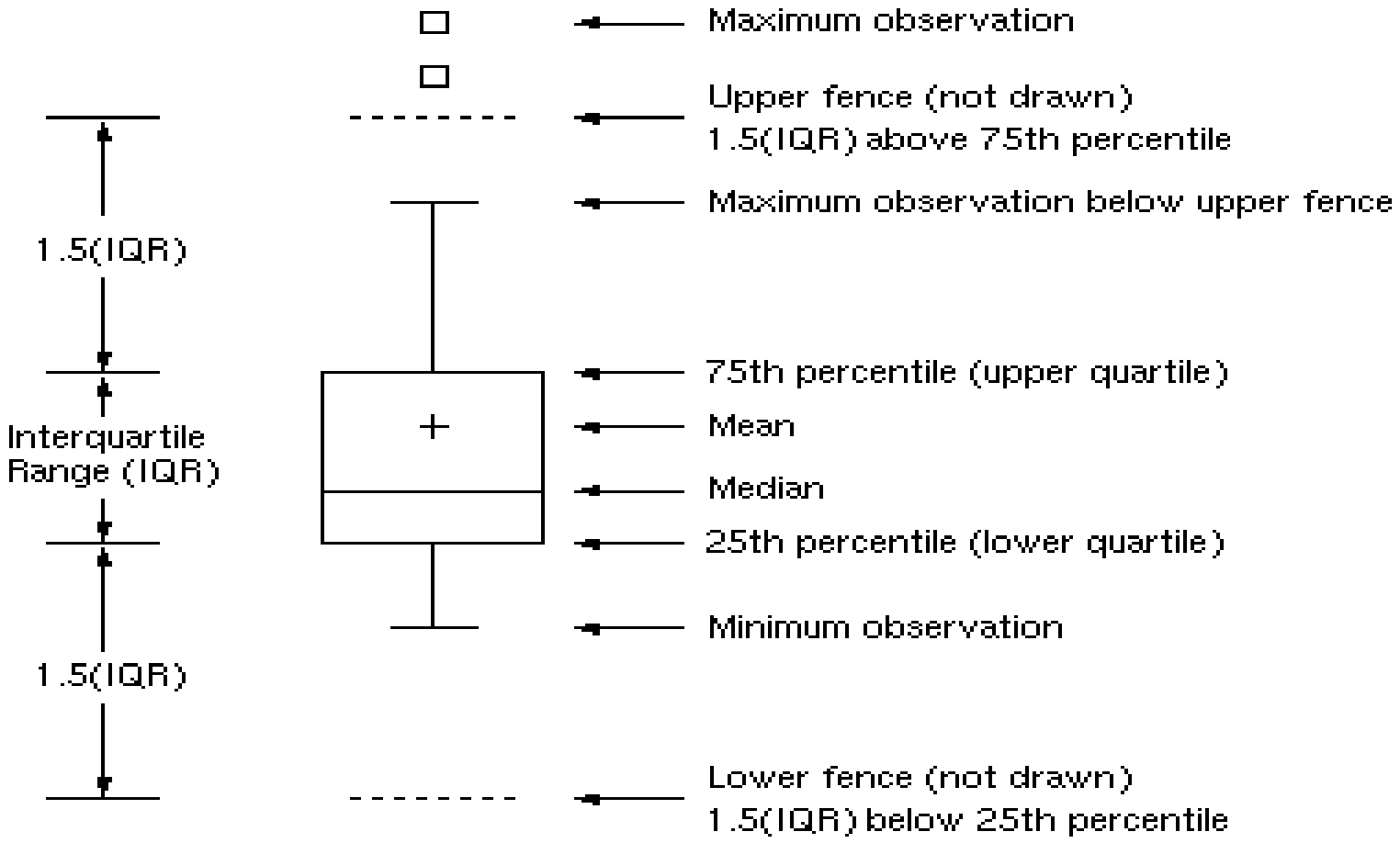
- The center of the data set;
- Where most of the data fall;
- The spread of the unquestionably ‘good’ data;
- Possible outliers;
- And more……
- Again, we introduce steps to draw boxplots but rely on R.

# A Boxplot from Texas House Data



# Schematics of Boxplot

- Schematic Box-and-Whiskers Plot



# Seven Steps for Boxplots

1. Construct a vertical scale, marked clearly, that covers at least of the data.
2. Find the median and the quartiles ($Q_1$ and $Q_3$).
3. Find the step size:
   - Step = $1.5 * (Q_3 - Q_1)$($Q_3 - Q_1$ is called **inter-quartile range**)
4. Find the inner fences, which define the bounds for questionably good data. **Upper Inner Fence (UIF)** and **Lower Inner Fence(LIF)** are given by: UIF = $Q_3$ + Step and LIF = $Q_1$ - Step.

# Seven Steps for Boxplots

5. Locate the most extreme data values on or within the inner fence, draw vertical lines at these points and then draw whiskers to connect these points.
6. Find the outer fences for discriminating between mild and extreme outliers. **Upper Outer Fence (UOF)** and **Lower Outer Fence(LOF)** are:

   UOF = $Q_3$ + 2 * Step

   LOF = $Q_1$ - 2 * Step

# Seven Steps for Boxplots

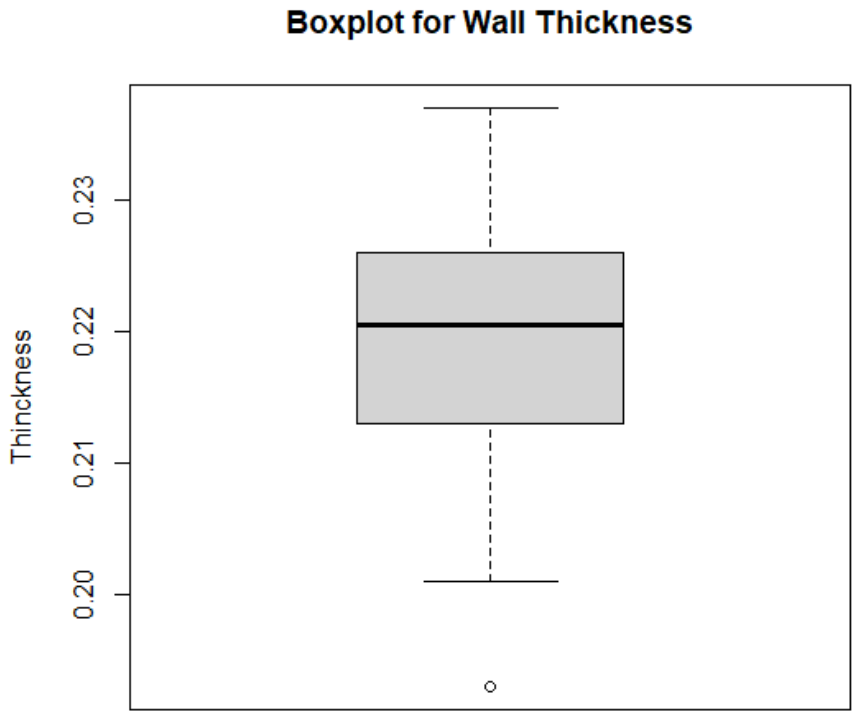
7. Mark possible outliers.
   - Use a ‘o’ to denote mild outliers, which are those data points between inner and the outer fences.
   - Use a ‘●’ to denote extreme outliers, which are those points on or beyond the outer fences.
   - You can use other symbols too.

# Example – Wall Thickness of Aircraft Parts

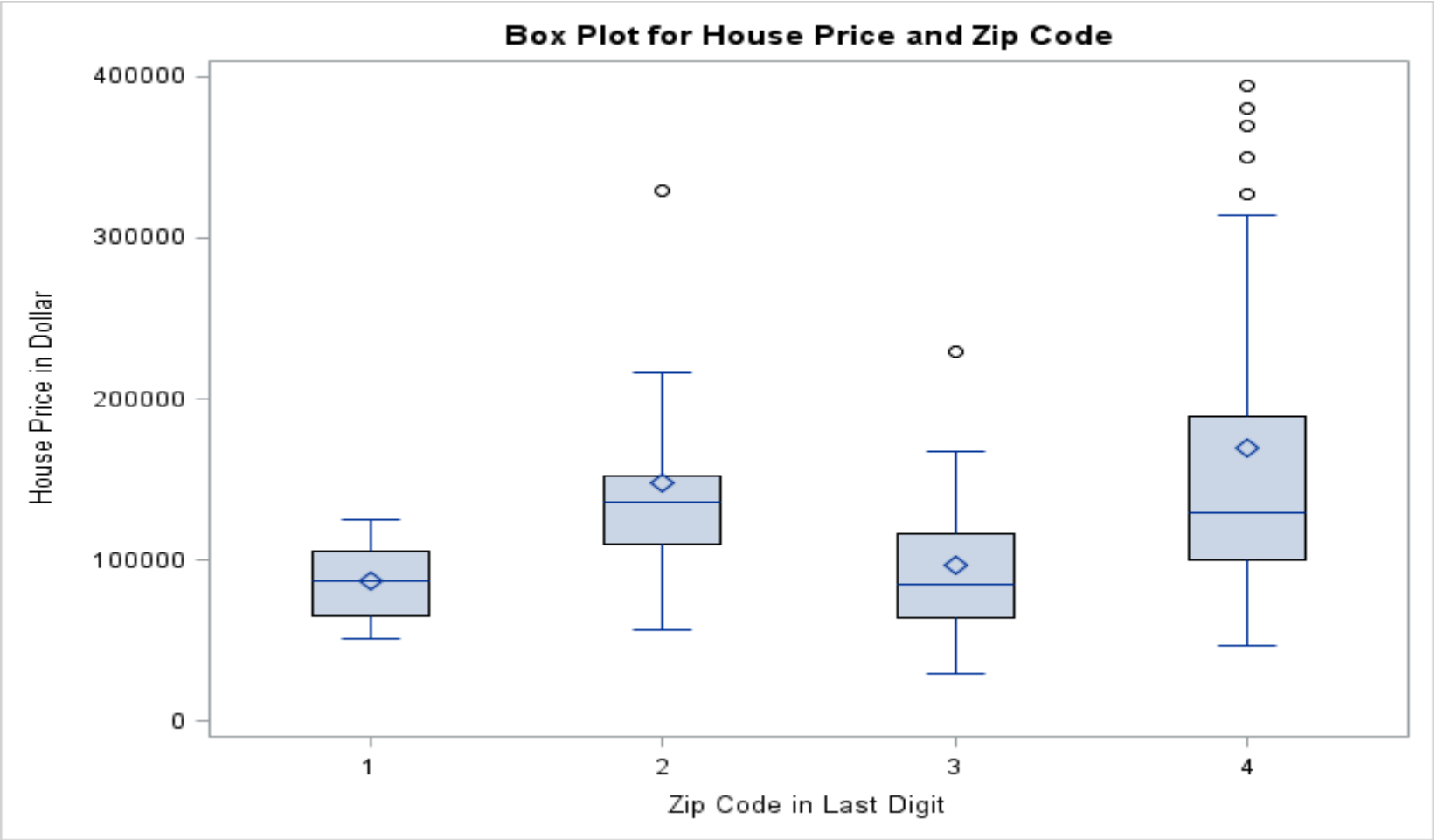
- **Example** – Eck Industries, Inc. Manufacturers cast aluminum cylinder heads that used for liquid-cooled aircraft engines. The thicknesses (in inches) for 18 cylinder heads are given below:

| | | | | | |
|---|---|---|---|---|---|
| 0.223 | 0.193 | 0.218 | 0.201 | 0.231 | 0.204 |
| 0.228 | 0.223 | 0.215 | 0.223 | 0.237 | 0.226 |
| 0.214 | 0.213 | 0.233 | 0.224 | 0.217 | 0.210 |

# Boxplot – Wall Thickness of Aircraft Parts



# Boxplot – Texas House Data



# Frequency Table for Two Variables

- A two-dimensional contingency table should be used for two qualitative/discrete variables.
- Use R function table(var1, var2) for frequency
  - A row is for a level of var1
  - A column is for a level of var 2
- Use R function prop.table(x, margin) for frequency
  - The input is the output of table()
  - margin: default - all data; = 1 – for rows; = 2 – for columns
  - margin of 1 or 2 is preferred

# Frequency Table for Two Variables

- **Interpretation**: Look at different patterns between rows/columns
- **Frequency Table**: table(texas$exter, texas$bed)

| Exterior | Number of Bedrooms | | | | |
|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** |
| **Brick** | 0 | 1 | 30 | 14 | 3 |
| **Frame** | 0 | 1 | 7 | 0 | 0 |
| **Others** | 1 | 1 | 9 | 2 | 0 |

- **Interpretation**: Brick houses seem to have greater number of bedrooms.

# Frequency Table for Two Variables

- **Interpretation**: Look at different patterns between rows/columns
- **Frequency Table**: prop.table(, margin = 1)

| Exterior | Number of Bedrooms | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Brick | 0.000 | 0.014 | 0.438 | 0.203 | 0.043 |
| Frame | 0.000 | 0.014 | 0.101 | 0.000 | 0.000 |
| Others | 0.014 | 0.014 | 0.130 | 0.029 | 0.000 |

- **Frequency Table**: prop.table(, margin = 2)

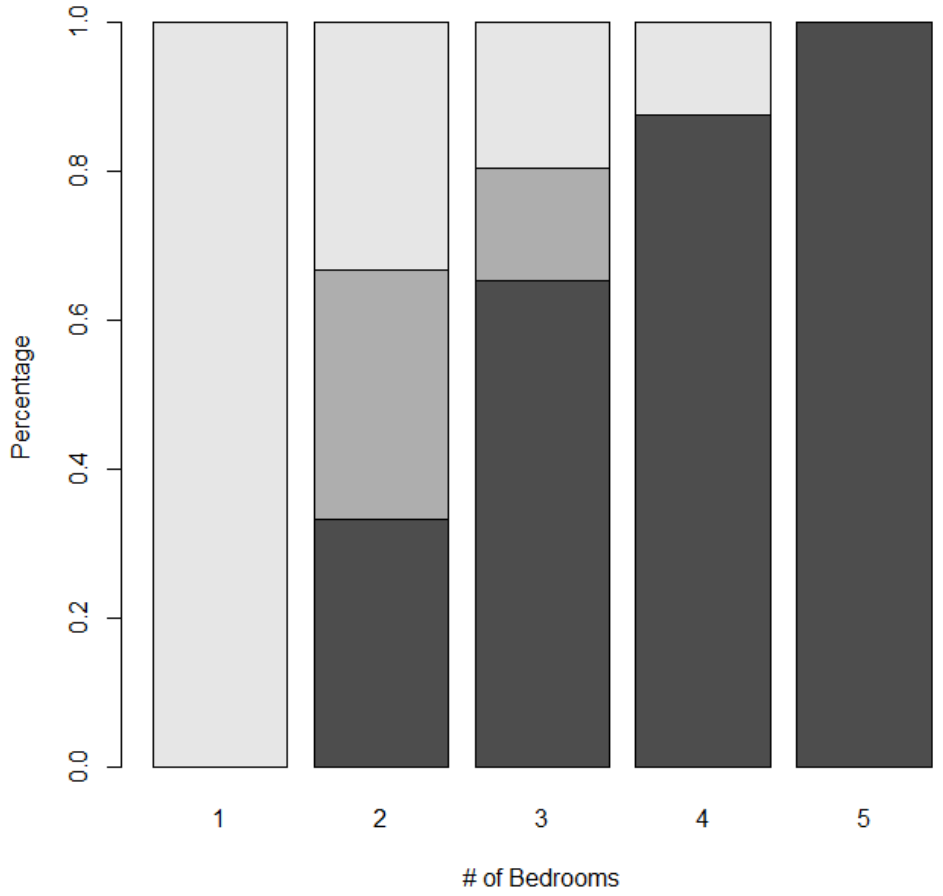
| Exterior | Number of Bedrooms | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Brick | 0.000 | 0.333 | 0.652 | 0.875 | 1.000 |
| Frame | 0.000 | 0.333 | 0.152 | 0.000 | 0.000 |
| Others | 1.000 | 0.333 | 0.197 | 0.125 | 0.000 |

# Bar Plot for Two Variables

- **Interpretation**: Look at different patterns between rows/columns
- Use R function barplot()
  - Input should be a two-dimensional table from prop.table()
  - y-axis: each level of the column variable
  - Bars are stacked according to levels of the row column
  - For prop.table(), margin = 2 should be used
- R program:

```
a <- table(texas$exter, texas$bed)
b <- prop.table(a, margin = 2)
barplot(b)
```

# Bar Plot for Two Variables

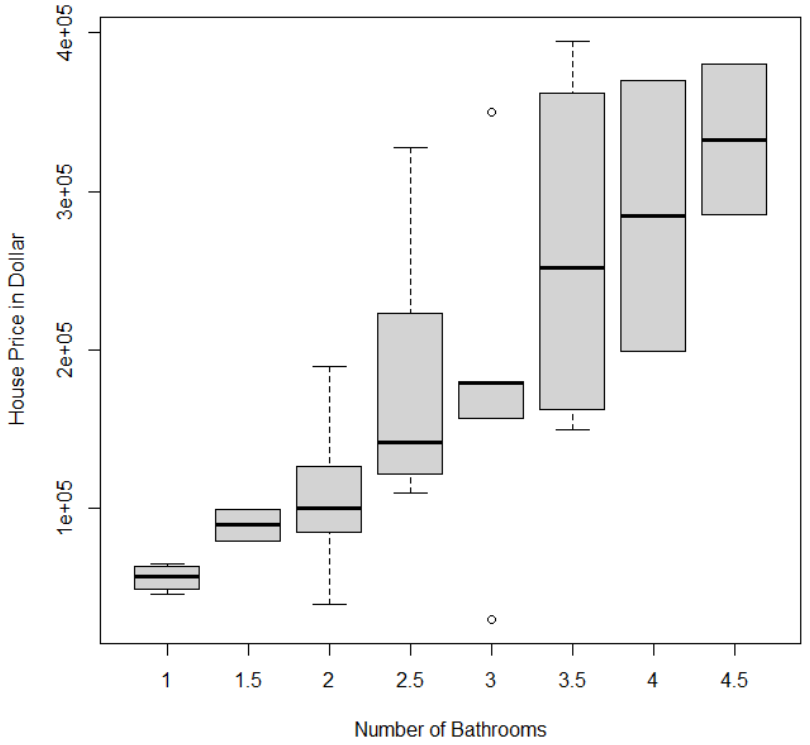


# Boxplots and Scatter Plots

- Boxplots for one qualitative/discrete variable and one quantitative variable
  - Use R function boxplot()
  - **Interpretation**: focus on difference/trend of means and difference of variability
- Scatter for two quantitative variables
  - Use R function plot()
  - **Interpretation**: focus on pattens/trends/variabilities

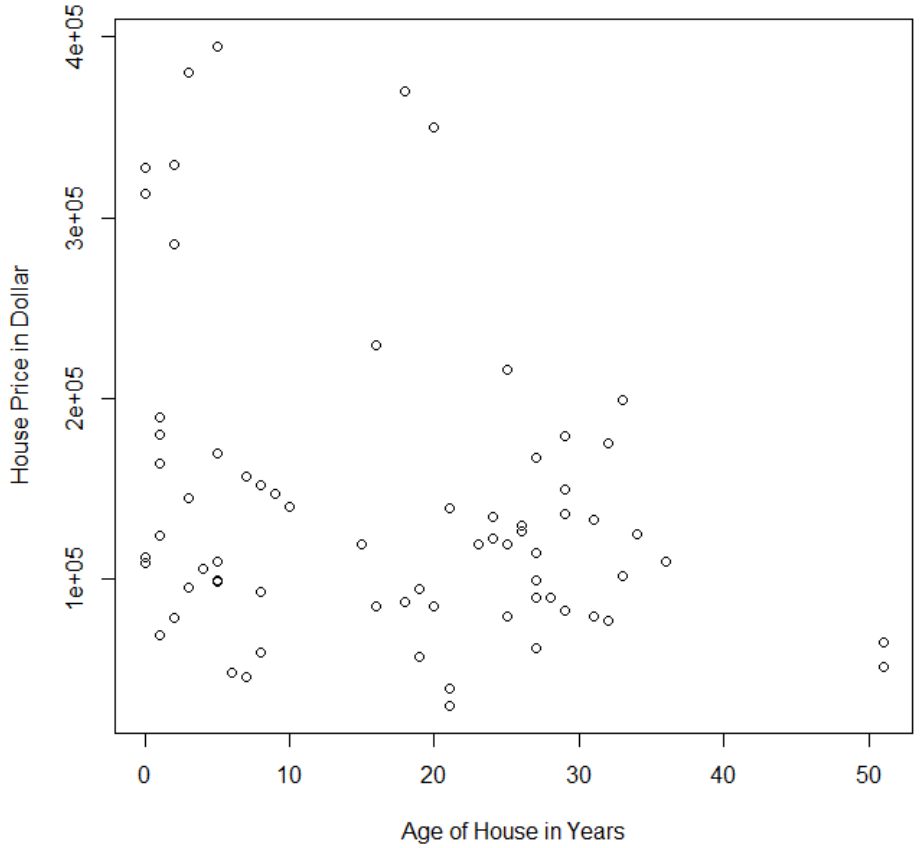
# Boxplots for Two Variables

- R code: boxplot(price ~ bath, data = texas)
  - First variable should be quantitative (y-axis)
  - Second variable should be qualitative or discrete (x-axis)
  - Use data argument for data frame used in the plot
  - **Interpretation**
    - Price increases with number of bathrooms.
    - The price for house with more than 3 bathrooms shows more variability.



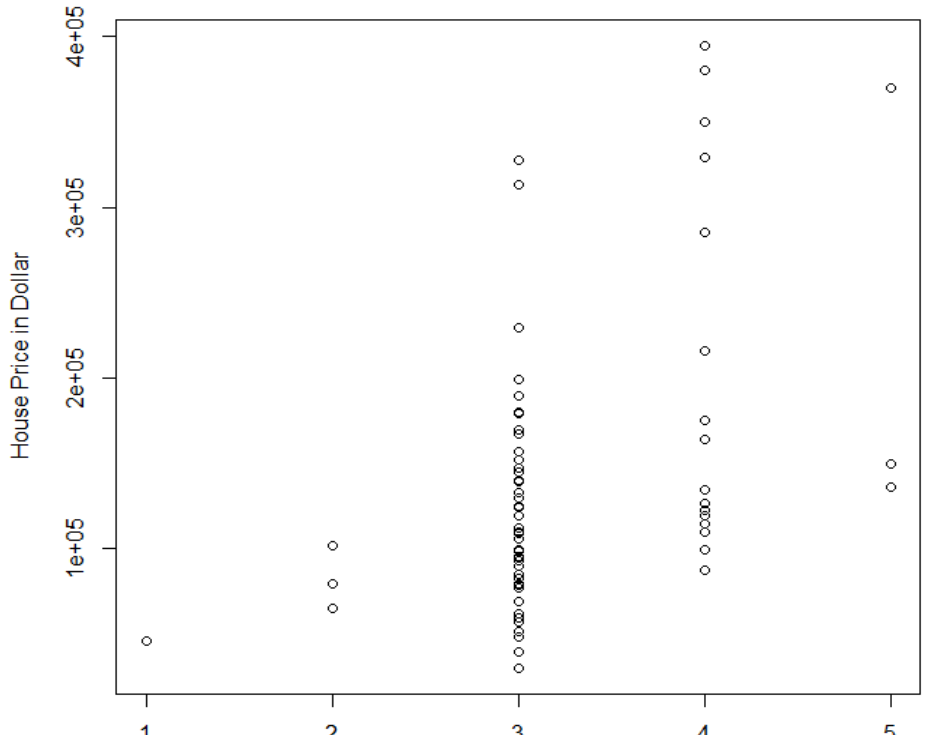
# Scatter Plots for Two Variables

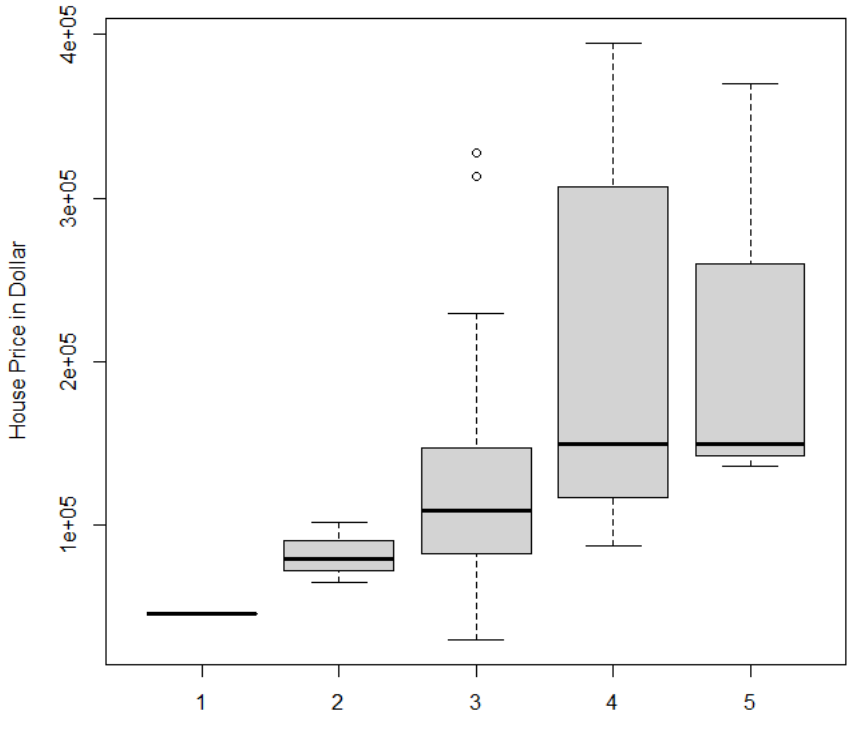
- R code: plot(price ~ age, data = texas)
  - First variable (y-axis) and second variable (x-axis) should be chosen carefully
  - Use data argument for data frame used in the plot
  - **Interpretation**
    - Some new houses have quite high prices
    - Prices for most houses are not related to their ages
    - Price for newer houses seem to have a wider range



# Scatter Plots for Two Variables

- R code: plot(price ~ bed, data = texas) (boxplot is better here)
- House price increases with the number of bedrooms.
- Price for houses with more bedroom seem to have a wider range



# Preview – Populations, Samples, Statistical Inference

- **Definition** – The **population** is the values of one or more variables for the entire collection of units relevant to a particular study
- **Population Parameters** – mean and variance, unknown, must be estimated from samples
- **Estimates** – the descriptive measures from samples, can reflect the population parameters, different from different sets of samples, how good an estimate is measured by "sampling error"
- **Statistics** – In this book, it is considered as the same as the estimate. In statistical theory, it refers to the function of a sample, which is either a random variable or random vector

# Data Collection

- **Goal** – make statements about population according to samples
- Radom sampling or some more advanced probability sampling is the appropriate way to collect data. In this book, we assume all samples are from “simple random sampling”
- **Definition** – The **simple random sampling** is a sampling scheme that each possible sample of the specified size has an equal chance of occurring
- Random sampling can be difficult to implement in practice
- Convenience samples are dangerous (Be careful)
- Sample size and power calculation

# Chapter Summary

- **Statistics, Data, Sample, Population**
- **Variable** – quantitative, qualitative
- **Variable** - nominal, ordinal, discrete, continuous
- **Table** – Frequency table for one or two variables
- **Statistics** – mean, variance, standard deviation, largest, smallest, median, quartile
- **Graphic** – boxplot, histogram, scatter plot, etc.

# Writing Report

- Use appropriate tables and figures to summarize data
- Do not directly copy tables or output from R output unless you are instructed to do so, do some edits (e.g., add descriptions, appropriate row and/or column names, effective digit)
- Discrete variables – report frequency and percentage
- Continuous variables – mean, variance or standard deviation