STAT 103
Statistical Thinking

Final Exam

Professor Mengshi Zhou

Final Exam

- PART I Producing Data: Chapter 1-Chapter 5
- PART II Summarizing Data: Chapter 10-Chapter 13
- PART III Regression and Chance: Chapter 14-Chapter 15; Chapter 17-Chapter 20
- PART IV Inference: Chapter 21-Chapter 22

Chapter 1 Individuals and Variables

Goal: Identify the individuals and the variables in a study(dataset)

Goal: Distinguish categorical variable and numerical variables

Individuals are objects described by the (people, animals, or things)

Variables any characteristics about individuals

Categorical variables place an individual in to one of serval groups (sex, major).

Numerical variables take numerical values for which arithmetic operations make sense (weight, height).

Chapter 1 Observational studies and Experiments

Goal: Know the definition and goal of observational studies and Experiments, and the difference between an observational study and an experiment

- **Observational studies** try to gather information without disturbing the scene they are observing.

- **Experiments** assignment treatment to individuals in order to see how they respond. The goal of an **experiment** is usually to learn whether some **treatment** actually causes a certain response.

Chapter 1 Population and Sample

Goal: Identify the population and sample in a study

Population: The entire set of possible individuals about which we want information

Sample: A subset of the population from which data are collected

Chapter 1 Sample Survey and Census

Goal: Know the definition of sample survey and census, determine if an observational study is a sample survey

- A **census** is an attempt to collect data from every member of the population.
- A **sample survey** is a collection of data from a subset of the population chosen by the researcher.

Chapter 2 Biased Samples

Goal: Be able to recognize bad sampling methods and know why they cause bias

- A **convenience sample** select of whichever individuals are easiest to reach
- A **convenience sample** is biased because it favors those who are easy to access y researchers

Chapter 2 Biased Samples

Goal: Be able to recognize bad sampling methods and know why they cause bias

- In a **voluntary response sample** individuals chooses themselves by responding to a general appeal (write-in or call-in opinion polls).
- A voluntary response sample is biased because people who feel strongly about a topic are more likely to respond to voice their feelings.

Chapter 2 Simple Random Sample

Goal: Know the definition of a simple random sample and SRS can avoid bias

- The deliberate use of chance in producing data is one of the big ideas of statistics. **Random samples** use chance to choose a sample, thus avoiding bias due to personal choice.
- The basic type of random sample is the **simple random sample**, which gives all samples of the same size the same chance to be the sample we actually choose.

Chapter 2 Simple Random Sample

Goal: Know the steps to generate a simple random sample

To select a simple random sample:

1. Label each population element with as few digits as possible, making sure each label is the same length.

2. Use the table or software to select random numbers.

Chapter 3 Proportion

Goal: Calculate sample proportion

Proportion

$$Proportion = \frac{Number\ in\ the\ category}{Total\ number}$$

The symbol for a sample proportion is $\hat{p}$ and is read as p-hat. The symbol for a population proportion is p.

Chapter 3 Parameter and Statistic

Goal: Identify of parameter and statistic in a study

Parameter: A number that describes a population. It is a fixed number, but in practice we don't know the actual value of this number.

Statistic: A number that describes a sample. This is a known value when we have taken a sample, but it can change from sample to sample. It is often used to estimate an unknown parameter.

Chapter 3 Bias and Variability

Goal: Understand Bias and Variability and know how to reduce Bias and Variation

Bias: When the design of a statistical study systematically favors certain outcomes. Can be reduced using simple random sample

Variability: Describes how spread out the values of the sample statistic are when we take many samples. Can be reduced using larger sample size

Chapter 4 Sampling Errors

- **Goal: Know the types of sampling errors**

- **Sampling errors** come from the act of choosing a sample. **Random sampling error, bad sampling methods** and **undercoverage** are common types of sampling error.
- **Undercoverage** occurs when some members of the population are left out of the **sampling frame**, the list from which the sample is actually chosen.

Chapter 4 Non-Sampling Errors

Goal: Know the types of nonsampling errors

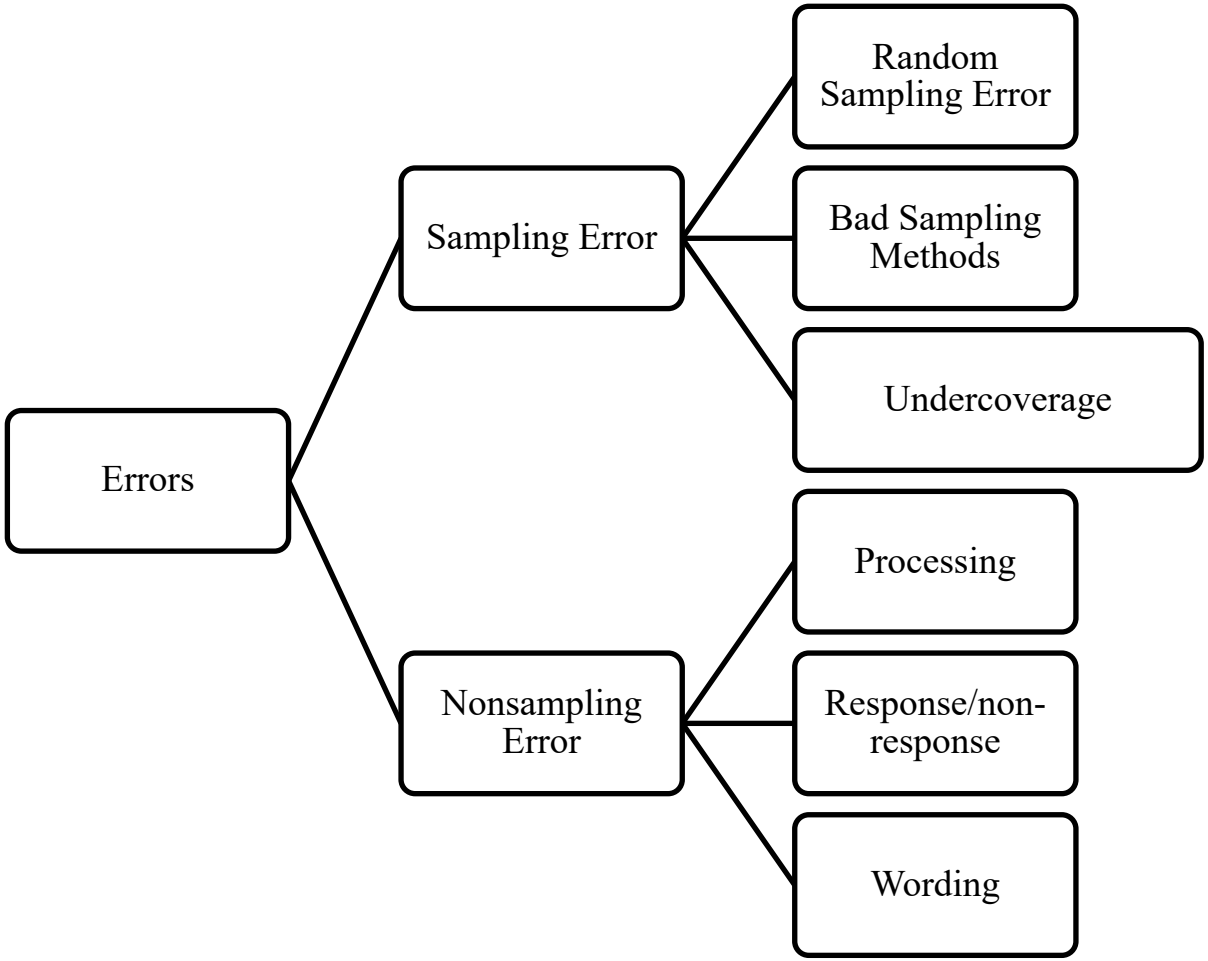
The most serious errors in most careful surveys, however, are **nonsampling errors**. These have nothing to do with choosing a sample—they are present even in a census.

The single biggest problem is **nonresponse**: subjects can't be contacted or refuse to answer.

Mistakes in handling the data (**processing errors**) and incorrect answers by respondents (**response errors**) are other examples of nonsampling errors.

Finally, the exact **wording of questions** has a big influence on the answers.

Chapter 4 Distinguish two types of errors



Margin of Error only covers random sampling error!

Chapter 4 Stratified Random Sample

Goal: Know what a stratified sample is and how to identify strata

Stratified random sample: A sample in which the sampling frame is first divided into various **strata** (groups). A simple random sample is then taken in each of these **strata**, with those selected combined to form the complete sample:

Chapter 5 response variables, explanatory variables, treatments

Goal: Identify the response variables, explanatory variables, and treatments in an experiment.

explanatory variable A variable that we think explains or causes changes in the response variable.

response variable A variable that measures an outcome or result of a study.

Treatments: Any specific experimental condition that is applied to the subjects. If an experiment has several explanatory variables, this is a combination of specific values of these variables.

Chapter 5 lurking variables and confounding

Goal: Be able to identify a lurking variable that is confounding a study

Lurking variable = has important effect on response variable but is NOT an explanatory variable (sometimes called a confounding variable or a third variable)

Two variables are **CONFOUNDED** when their effects on a response variable cannot be distinguished from each other.

Chapter 5 Randomized Comparative Experiment

Goal: Know the definition of a Randomized Comparative Experiment

Randomization--randomly assigning cases to different levels of the explanatory variable (e.g., different treatment groups).

An experiment that involves randomization may be referred to as a **randomized comparative experiment.**

Chapter 5 Logics for good experiments

Control Group--A group receive no treatment or a placebo

Placebo Group--A group that receives placebo (e.g., a sugar pill in a medication study)

If we want to compare multiple treatments, we do not need placebo group

Chapter 5 Logics for good experiments

Goal: Fully understand the logic of experimental design

Identify control groups, randomization, placebos, blinding, and use enough subjects in experiments and explain why each is used.

Chapter 5 Logics for good experiments

Control the effects of lurking variables on the response, most simply by comparing two or more treatments.

Randomize – use impersonal chance to assign subjects to treatments.

Double-Blind –Research study in which neither the participants nor the researchers interacting with them know which cases have been assigned to which treatment groups

Use enough subjects in each group to reduce chance variation in the results.

Make sure experimental units **represent** the others not in the experiments

Final Exam

- PART I Producing Data: Chapter 1-Chapter 5
- PART II Summarizing Data: Chapter 10-Chapter 13
- PART III Regression and Chance: Chapter 14-Chapter 15; Chapter 17-Chapter 20
- PART IV Inference: Chapter 21-Chapter 22

Chapter 10 Variable Type

- A **categorical variable** places individuals into categories

- A **quantitative variable** takes on numeric values for which arithmetic operations make sense

Chapter 10 Distribution

- We can use frequency tables to summarize data

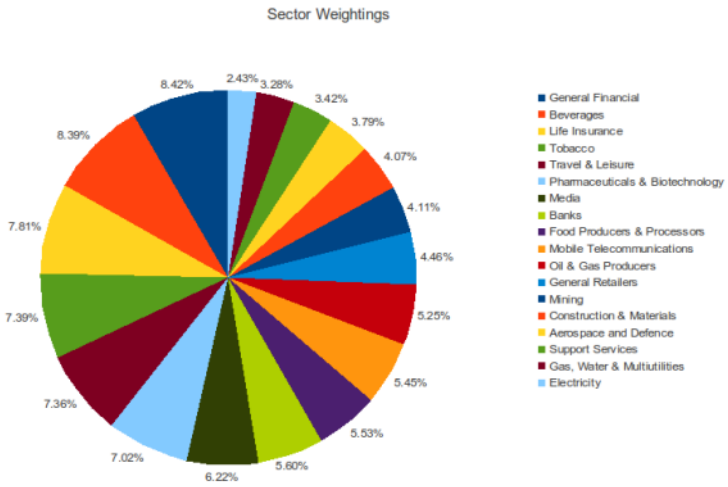
- The distribution of a variable:

Tells what values the variable can take and How often it takes those values.

Position	Number of Players	Rates
Catcher	34	7.7
First Baseman	19	4.3
Outfielder	83	18.9
Pitcher	215	49.0
Second Baseman	24	5.5
Shortstop	41	9.3
Third Baseman	23	5.2

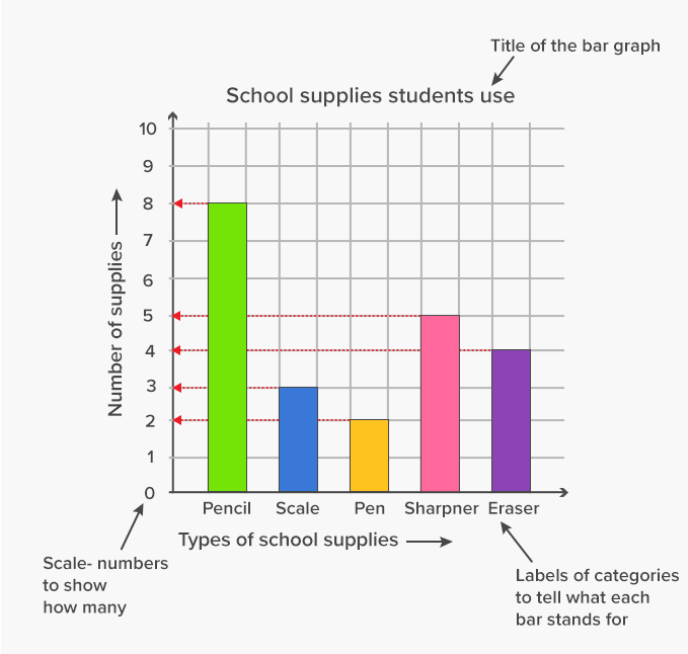
Chapter 10 Bar graph and pie chart

- The wedges of the pie chart correspond to the rate of each value
- A pie chart can only be used when it represents all the parts of one whole
- We cannot use pie charts to describe the counts.



Chapter 10 Bar graph and pie chart

- A bar graph is more versatile than a pie chart
- One axis represents all the possible value of a categorical variable
- The rates or counts will be displayed on the other axis and cover the whole range of possible values of the numbers counted for each category.



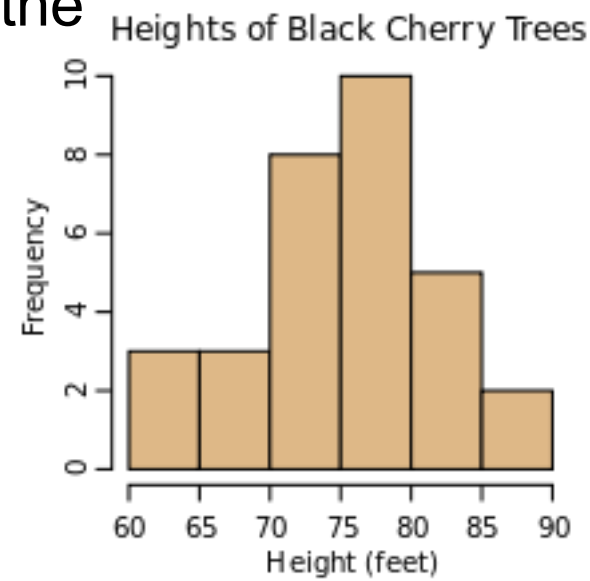
Chapter 11 Histograms

- Quantitative variables can be illustrated with histograms

- A histogram divides the range of data into classes of equal widths and then counts the number of observations in each class

  – The classes of a histogram

  – Must be of equal width

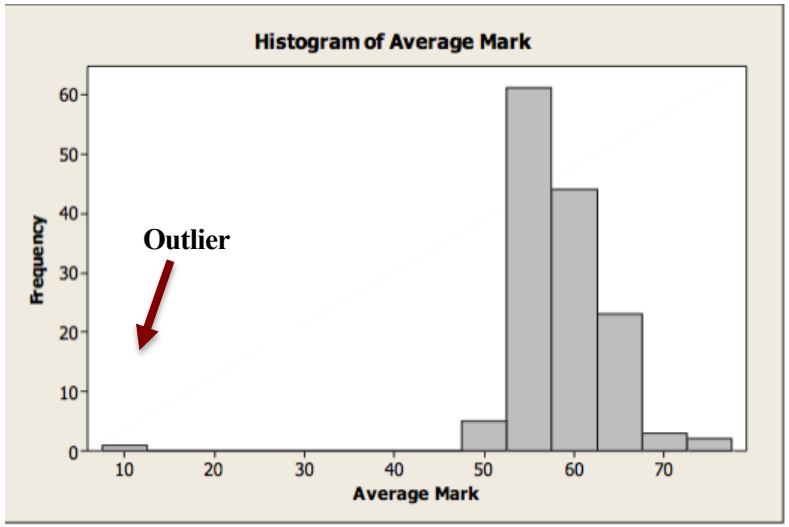


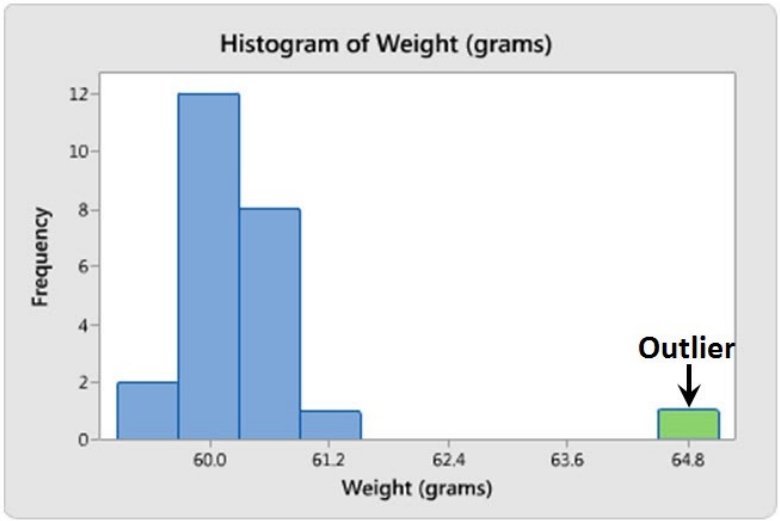
Chapter 11 Histograms

- In a histogram, look for:
  - Skewness
  - The number of peaks
  - Any outliers that fall outside the pattern of the graph

Outliers

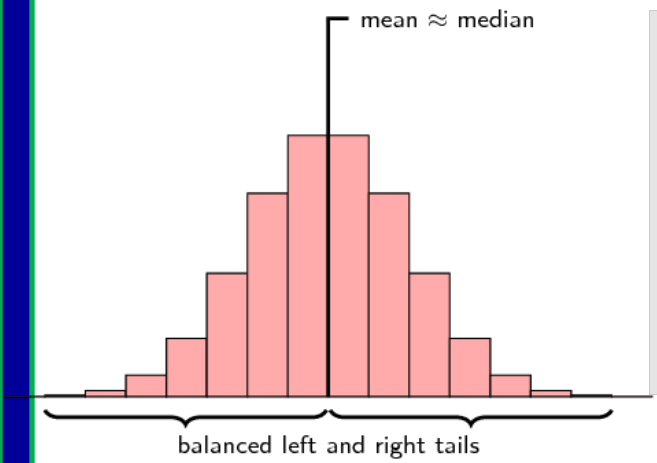
- Any data value away from the rest of the data a striking deviation

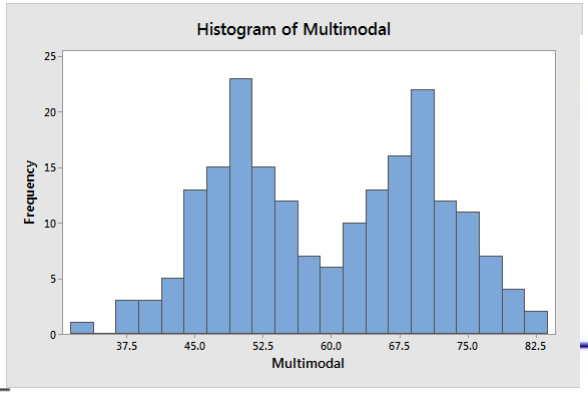


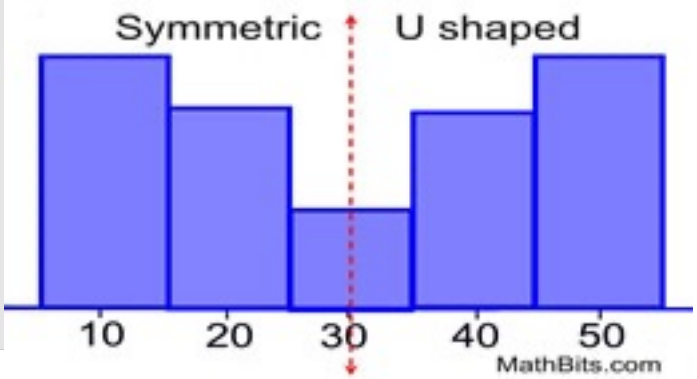


Shape: Symmetric or Skewed?

- Symmetric: A distribution is set to be symmetrical if it can be divided into two equal sizes of the same shape

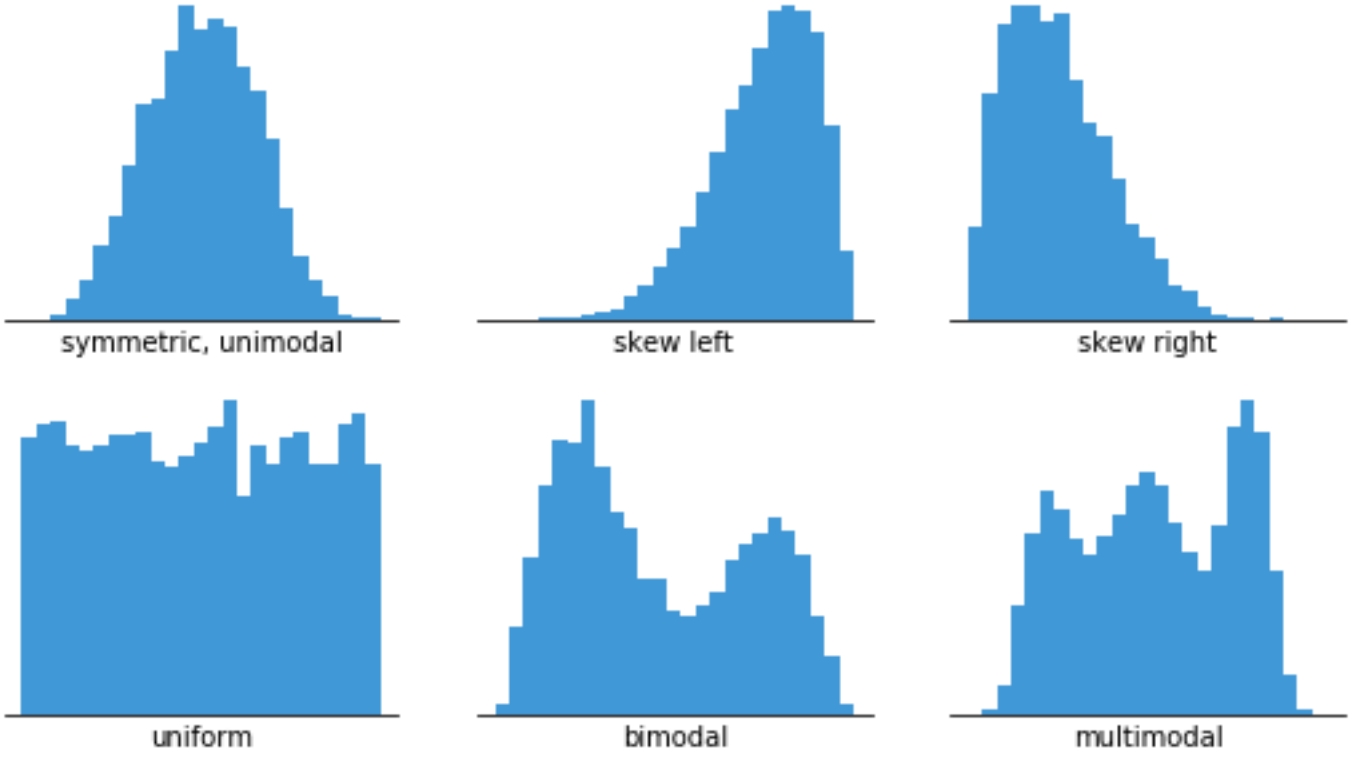






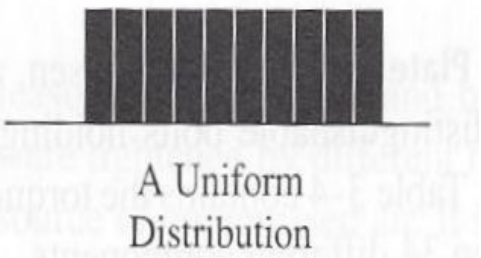
Shape: Symmetric or Skewed?

- Skewed: a skewed distribution refers to asymmetry distribution

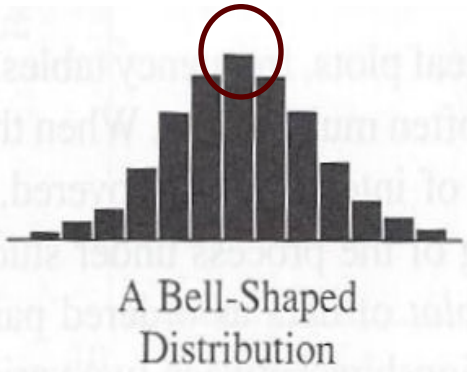


Shape: Peaks

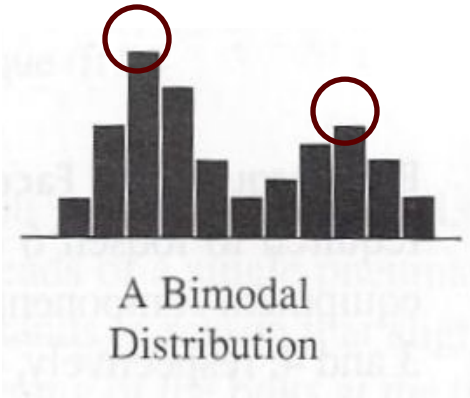
- Peaks occur when that data value is greater than its neighboring data points (on the left and right sides)



No peak-uniform



One peak-unimodal



Two peaks-bimodal

Chapter 12 Measure the Center

- The **mean** (the numerical average): The mean is found by obtaining a sum of all the observations and dividing by the sample size (n):

$$\overline{X} = \frac{\text{sum of observations}}{n}$$

- The **median** (the 50th percentile): The median is the middle value of a sample when the observations are sorted from smallest to largest. 50% larger than the median, 50% lower than the median.

median

- Half observations are less than median, the other half are larger than median
- The procedure of find **median (M)**
  - Order n observations, smallest to largest
  - Find the value that is in **location** $\frac{n+1}{2}$
    - If n is odd, median is the value in location $\frac{n+1}{2}$
    - If n is even, median is the average of the values in locations before and after $\frac{n+1}{2}$

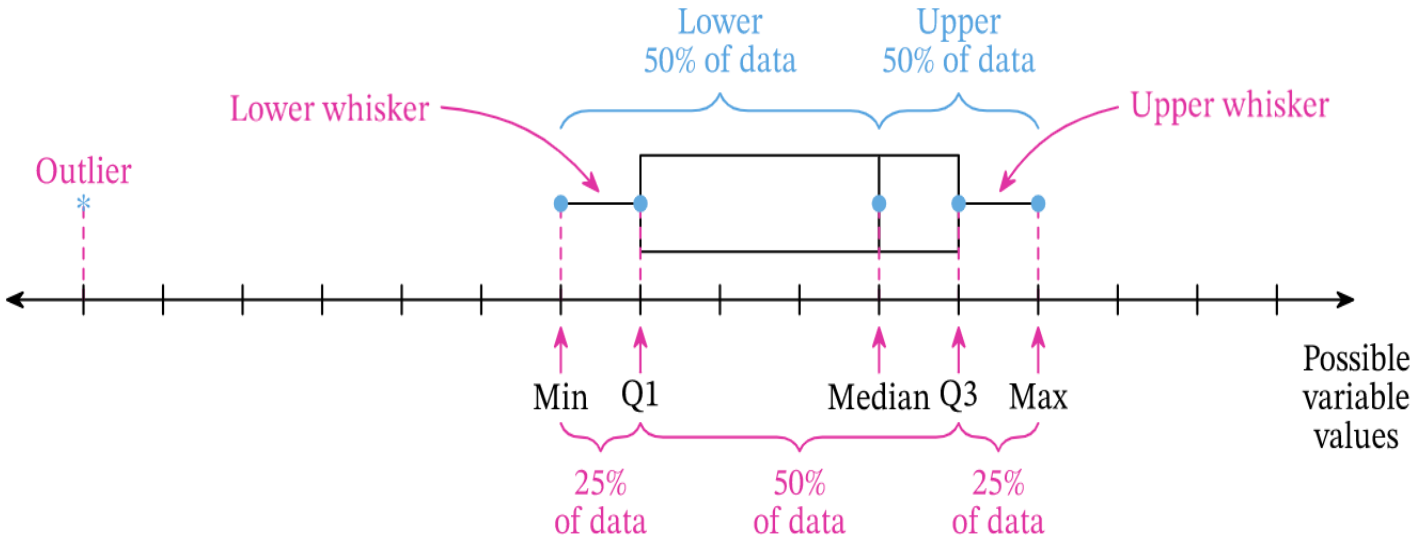
Chapter 12 Measure the Spread

- The **standard deviation is the average distance from the mean**
  - As the observations become more spread from their mean, the standard deviation gets larger.
  - It is used to describe the spread only when the mean is used to describe the center
  - It is equal to zero only when there is no spread at all

Chapter 12 Measure the Spread

- The first quartile (Q1) locates the middle of the lower half of the data. So, 25% of the data sit below Q1 and 75% of the data sit above Q1.
- The third quartile (Q3) locates the middle of the upper half of the data. So, 75% of the data sit below Q3 and 25% of the data sit above Q3.
- The interquartile range (IQR) = Q3 – Q1

Chapter 12 Boxplot and five-number summary



Chapter 12 Choose the right number of summarizing the distribution

- If a **sample has outliers and/or skewness**, resistant measures (median and IQR) are preferred over sensitive measures. This is because sensitive measures tend to overreact to the presence of outliers.

- If a sample is **reasonably symmetric with no outliers**, sensitive measures (mean and standard deviation) should be used. It is always better to use all of the observations in the sample when there are no problems with skewness and/or outliers.

Chapter 12 Skewness

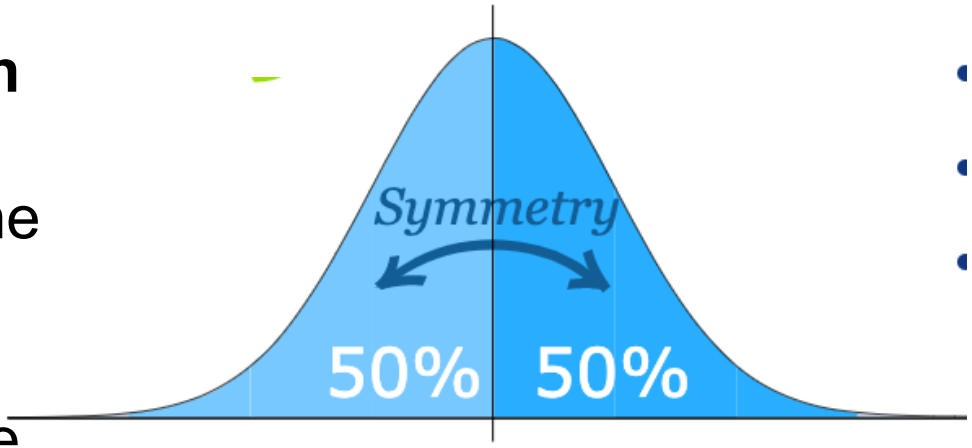
Histogram Shape	Compare Two Measures of Centers
If symmetric	mean and median are approximately equal
If right skewed	mean is greater than the median
If left skewed	mean is less than median

Chapter 13

- Density Curve
- Normal Distribution
- 68-95-99.7 rule
- Standard score = $\frac{(observation - mean)}{(standard\ deviation)}$
- Percentiles

Summary of Normal Distribution

- The normal distribution is **unimodal**
- The normal distribution is **symmetric about its mean**
- The curve is on or above the horizontal axis.
- The **mean** determine where the data tends to cluster
- the **standard deviation** determine how spread the distribution will be

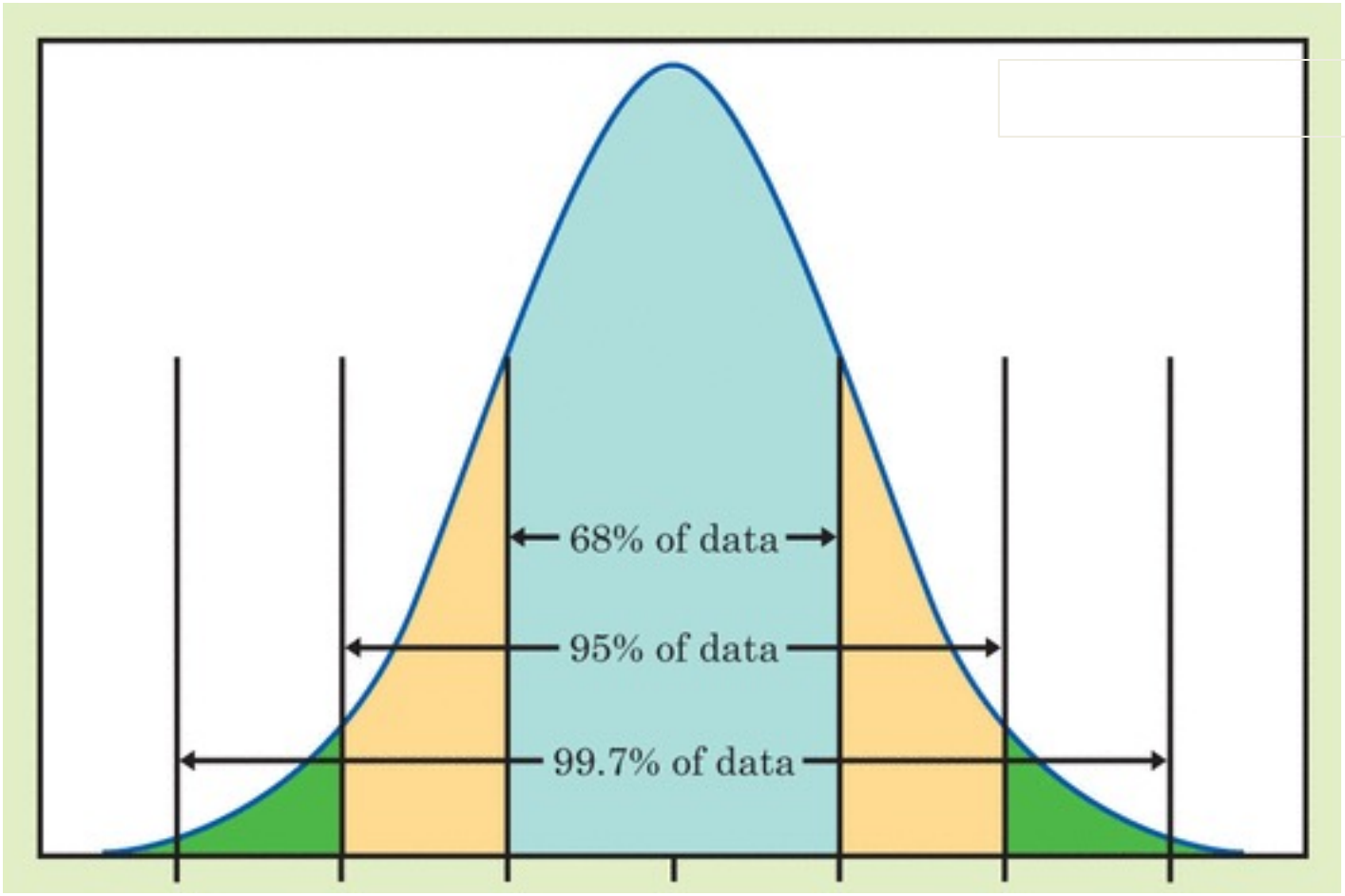


What's so special about being "normal"?

- In any Normal distribution, approximately
  - 68% of all observations fall within 1 standard deviation from the mean *(mean – 1sd to mean + 1sd)*
  - 95% of all observations fall within 2 standard deviations from the mean *(mean – 2sd to mean + 2sd)*
  - 99.7% of all observations fall within 3 standard deviations from the mean *(mean – 3sd to mean + 3sd)*

*Only **approximately** true for any set of data*
(No set of data is exactly described by a smooth curve)

68-95-99.7 Rule



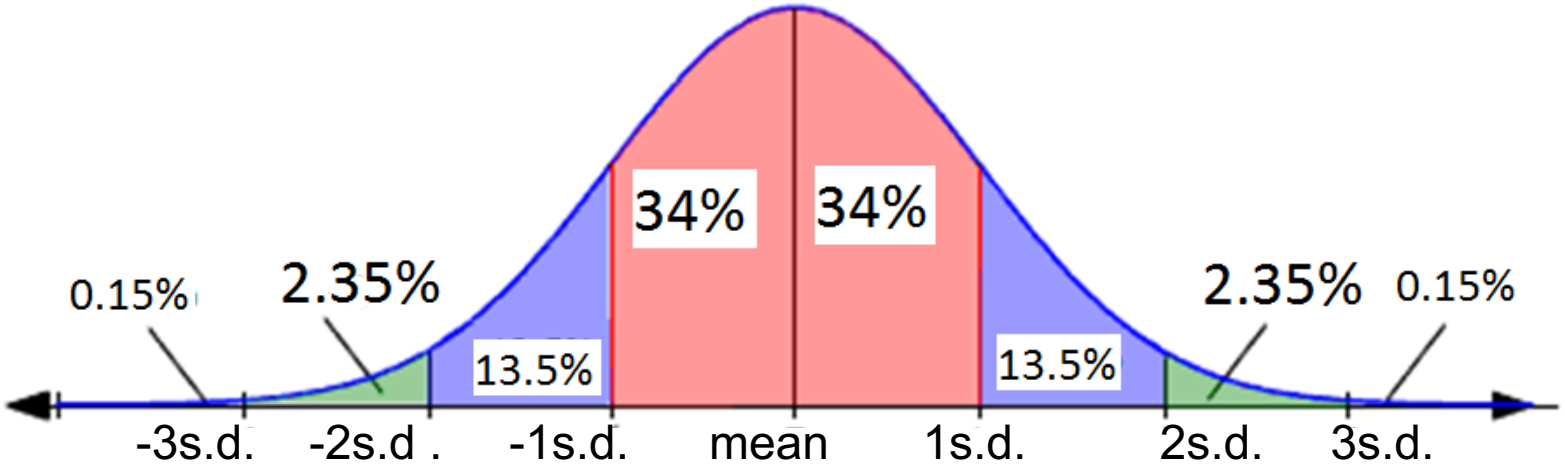
Updated-68-95-99.7 rule

- Because the normal distribution is “symmetric”, we can find more information
- Examples:
  - 34% of all observations fall between mean to mean + 1 standard deviation
  - 2.35%+0.15% = 2.50% of all observations are larger than mean + 2 standard deviation
  - How many observations are smaller than mean - 2 standard deviation?

2.35%+0.15% = 2.50%

How many observations are larger than mean + 1 standard deviation?

13.5% + 2.25% + 0.15% = 16%



Standard Scores

- We measure the distance an observation is away from the mean in standard deviation units.
  - *"this observation is 1 standard deviation above the mean"*
  - *"it falls 1.7 standard deviations below the mean"*

- The standardized score measure how many standard deviation a data point is above/below the mean
  - **Standard score** = $\frac{(observation - mean)}{(standard\ deviation)}$

- Most often called a **z-score**
- Use z-score to compare values, find percentages or percentiles
- The larger the z-score, the less likely for that value to occur

Percentiles

- Percentile: the n^{th} percentile is a value such that n percent of the observations lie below it (and the rest lie above it)
- If a person is in the 99^{th} percentile of height, this people is taller than 99% of the rest of the population
- If a number is in the 50^{th} percentile of height, this number is larger than 50 % of the observations, and smaller than 50 % of the rest of the observations
- If a number is in the 75^{th} percentile of height, this number is larger than 75 % of the observations, and smaller than 25 % of the rest of the observations

Final Exam

- PART I Producing Data: Chapter 1-Chapter 5
- PART II Summarizing Data: Chapter 10-Chapter 13
- PART III Regression and Chance: Chapter 14-Chapter 15; Chapter 17-Chapter 20
- PART IV Inference: Chapter 21-Chapter 22

Chapter 14: Scatterplot

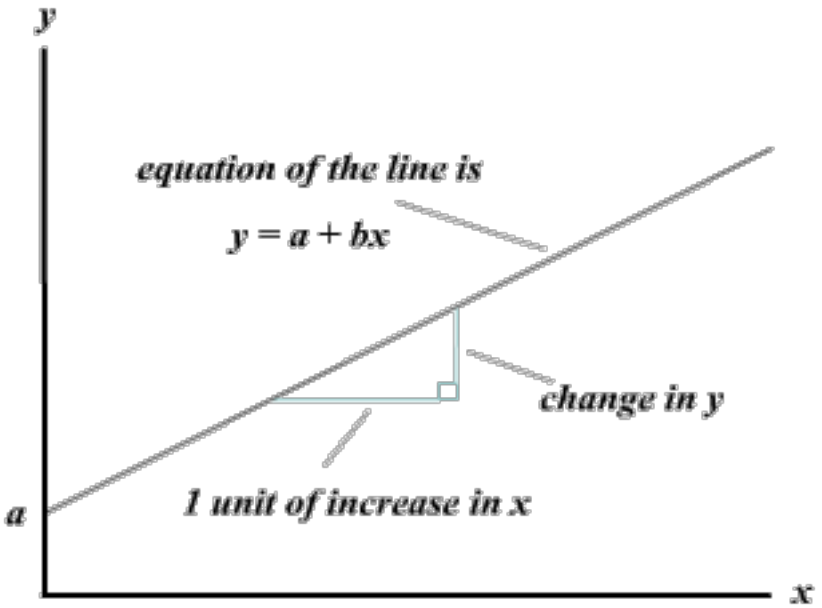
- How to read the scatter plot
- If there is a clear **direction**, is it positive (the scatterplot slopes upward from left to right) or negative (the plot slopes downward)?
- Is the **Form** straight (linear) or curved (non-linear)?
- **Strength:** How closely do the points follow the form (shape)
  - Strong, Moderate, Weak
- **Outlier:** any point(s) that don't fit the form or are far away from the rest of the points

Chapter 14: correlation *r*

- A positive correlation indicates a **positive linear** association. The strength of the positive linear association increases as the correlation becomes closer to +1.
- A negative correlation indicates a **negative linear association**. The strength of the negative linear association increases as the correlation becomes closer to -1.
- A correlation of **either +1 or -1** indicates a perfect linear relationship. This is hard to find with real data.
- Strong r>=0.8 or r=<-0.8; Moderate: 0.5<=r<0.8 or -0.8<r=<-0.5; Weak 0<r<0.5 or -0.5<r<0

Chapter 15: Regression Line

- Understanding the line
  - Equation of a line: $y = a + bx$
  - **b** is the **slope**
    - How much y changes on average when x increases by one unit
    - When b is **positive** there is a positive linear association,
    - when b is **negative** there is a negative linear association
  - **a** is the **y-intercept**
    - The value of y when x is zero

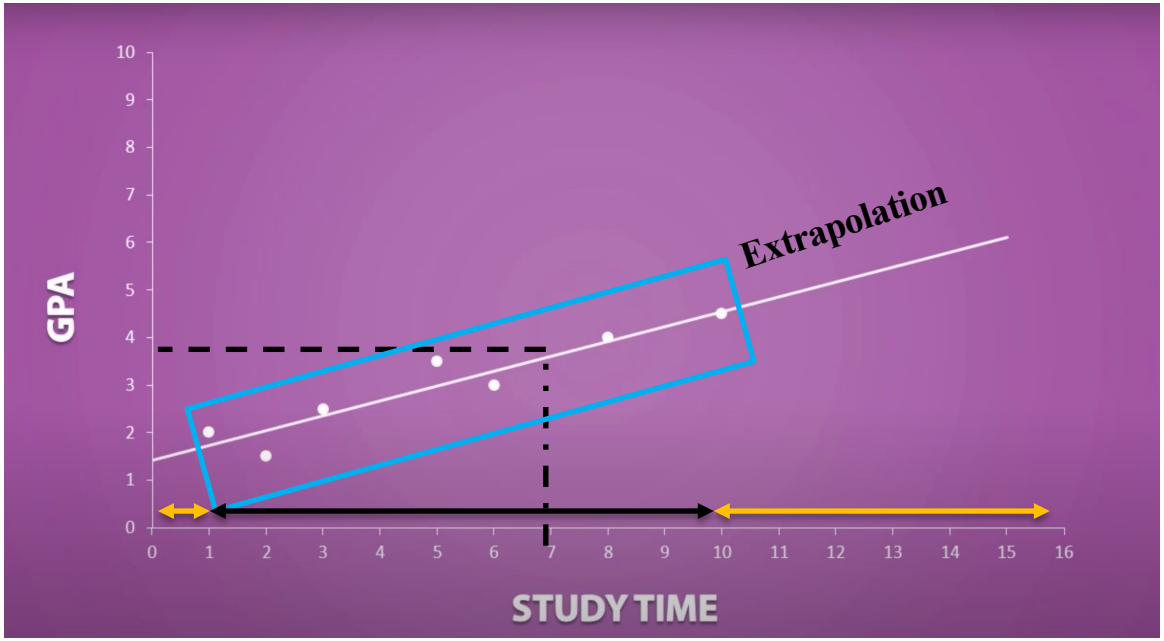


Regression Line Issues

- Only for linear relationships
- Outliers/influential points
- r^2
- Extrapolation

Extrapolation

Extrapolation: Make predictions outside the range of data



Coefficient of Determination, r^2

- r^2 (or R-squared) is a measurement of how well the regression line fits the data
  - r^2 = **r** * **r** (square of correlation) (e.g. r = 0.5, r^2 = 0.5*0.5 = 0.25)

r (correlation)	r^2 = r * r
Has value between -1 and 1	Has value between 0 and 1
Measures the linear relationships between two numerical variables with respect to directions and strength	r-squared also tells us the percentage of variation in Y that is accounted for by its regression on X

Correlation and Causation

- In an observational study, A strong correlation between two variables is not always evidence that changes in one variable cause changes in the other.

CHAPTER 17

Probability and Random:

- Some things in the world, both natural and of human design, are **random**. That is, their outcomes have a clear pattern in very many repetitions even though the outcome of any one trial is unpredictable.

- **Probability** describes the long-term regularity of random phenomena. The probability of an outcome is the proportion of very many repetitions on which that outcome occurs.

- $\text{Probability} = \frac{\#of\ favorable\ outcomes}{\#All\ possible\ outcomes}$

- A probability is a number between 0 (the outcome never occurs) and 1 (always occurs).

Law of Averages

Law of Averages (law of large numbers):

1. Mean or proportions are likely to be more stable when there are more trials;
2. while sums or counts are likely to be more variable.
3. This does not happen by compensation for a bad run of luck since independent trials have no memory.

CHAPTER 18

Probability model

- A **probability model** describes a random phenomenon by telling what outcomes are possible and how to assign probabilities to them.
- We sometimes call an outcome or a collection of outcomes an **event**. (question 1)

Status	No High School	H.S. or equivalent	Associate	Bachelor	Post-Bachelor
Probability	0.12	0.45	0.21	0.18	0.04

Probability Model Rules

1. A probability must Between 0 and 1
2. The sum of probabilities of all possible outcomes must have probability 1.
3. The probability that an event does not occur is 1 minus the probability that the event does occur.
4. If events have no common outcomes, the P(event) is the sum of the Probability of each outcome in the event :

   P(event A OR event B) = P(event A) + P(event B)

Sampling Distribution

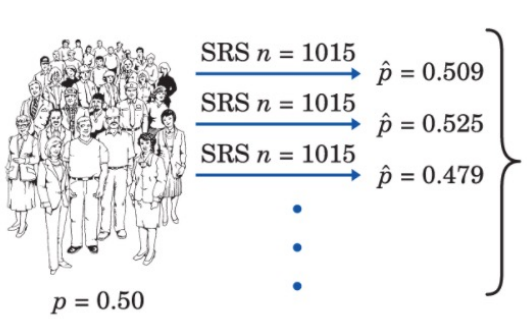
•Repeated Random Samples

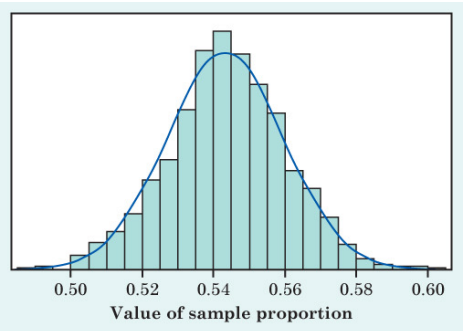
•In the long run, this distribution of statistics is well described by the normal curve.

•The mean of the normal curve is the value parameter

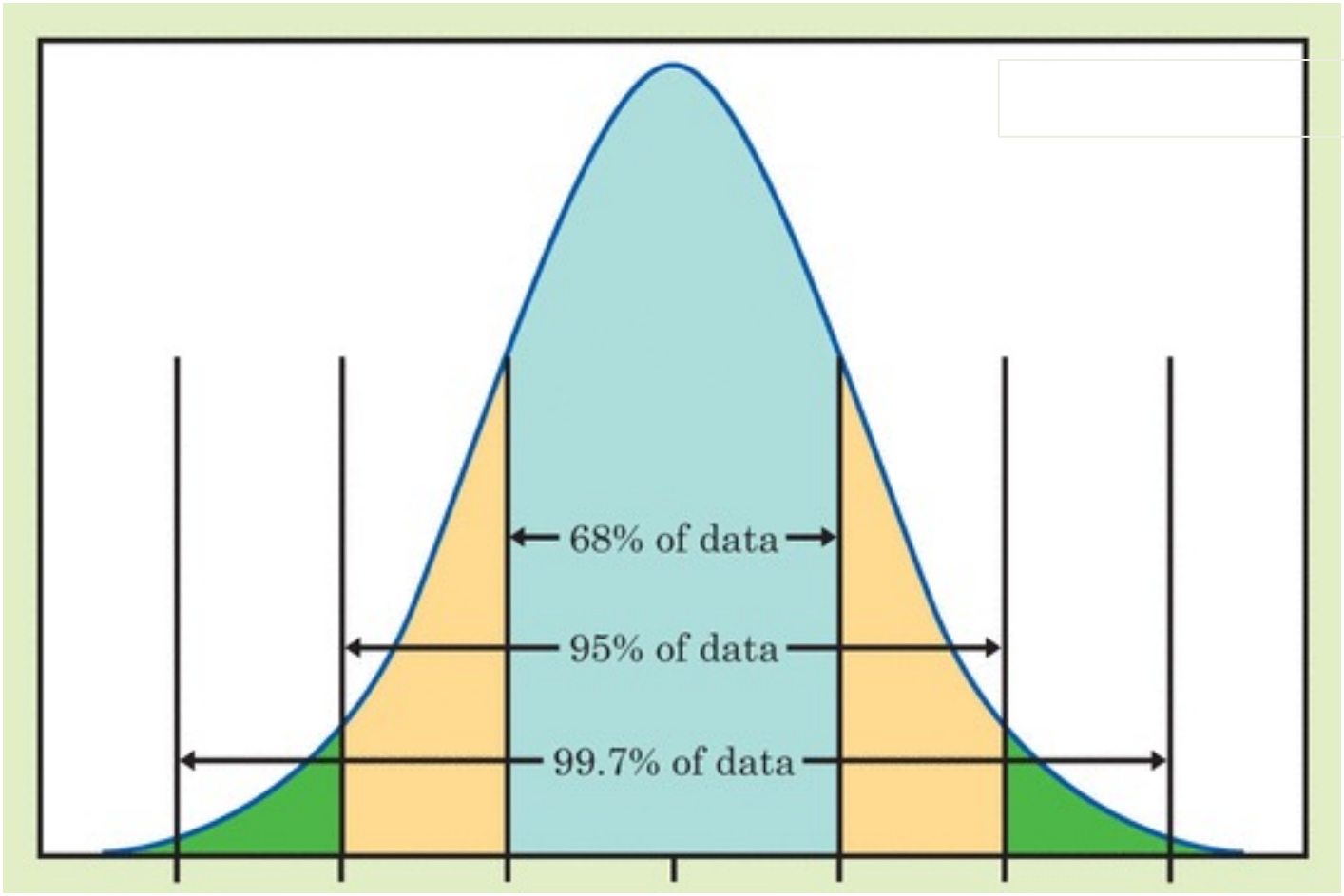
•The total probability is 1 because the total area under the curve is 1.

•We can use 68-95-99.7 rules!





68-95-99.7 Rule



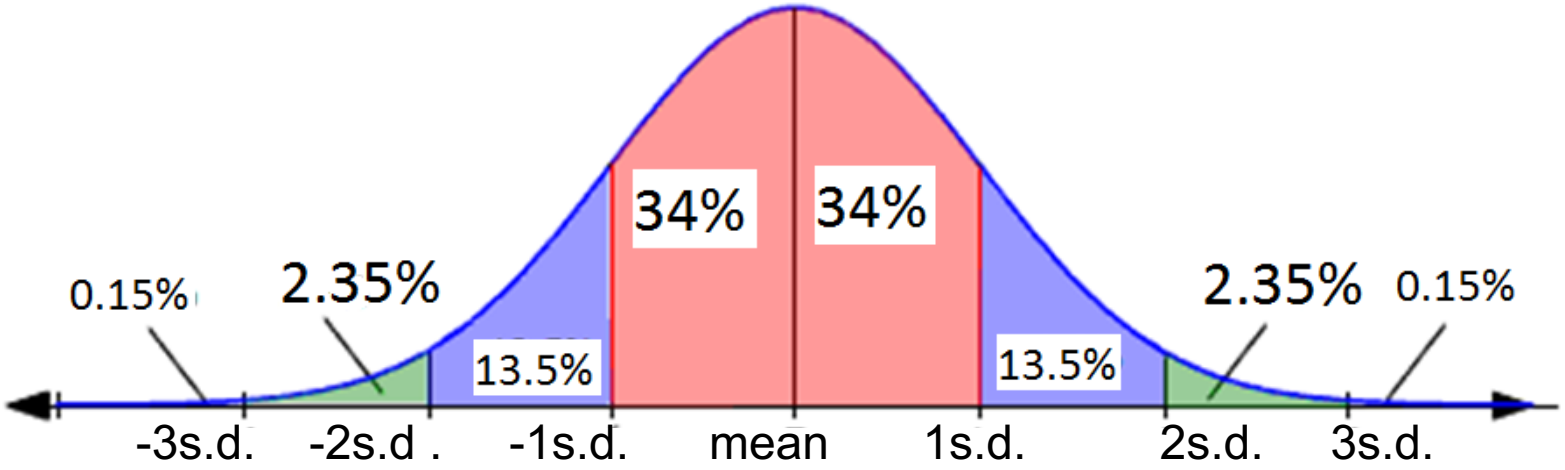
Updated-68-95-99.7 rule

- Because the normal distribution is “symmetric”, we can find more information
- Examples:
  - 34% of all observations fall between mean to mean + 1 standard deviation
  - 2.35%+0.15% = 2.50% of all observations are larger than mean + 2 standard deviation
  - How many observations are smaller than mean - 2 standard deviation?

2.35%+0.15% = 2.50%

How many observations are larger than mean + 1 standard deviation?

13.5% + 2.25% + 0.15% = 16%



Final Exam

- PART I Producing Data: Chapter 1-Chapter 5
- PART II Summarizing Data: Chapter 10-Chapter 13
- PART III Regression and Chance: Chapter 14-Chapter 15; Chapter 17-Chapter 18
- PART IV Inference: Chapter 21-Chapter 22

Statistics and Parameter

- Use sample proportions, $\hat{p}$ to estimate population proportions, p

Known-Statistic	Unknown-Parameter
Sample proportion, $\hat{p}$	Population proportion, p

Chapter 21 Confidence Intervals

	Population Proportion p
1. statistic	Sample Proportion $\hat{p}$
2. Standard error	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
3. Margin of Error	$z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
4. Confidence interval	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
5. Explain	We are …. confident that Population Proportion p is between $\hat{p} - z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ to $\hat{p} + z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Critical Values

TABLE 21.1 Critical values of the Normal distributions

Confidence level C	Critical value z^*	Confidence level C	Critical value z^*
50%	0.67	90%	1.64
60%	0.84	95%	1.96
70%	1.04	99%	2.58
80%	1.28	99.9%	3.29

Chapter 22 Hypothesis Tests

Tests of significance (Hypothesis test)

1. Hypothesize

   State the null (H_0) and alternative hypotheses (H_A)

2. collect data and figure out distribution assuming H_0 is true

   Normal, Middle (Mean), Standard deviation, A test statistic

	Population Proportion p
Middle (Mean)	Population Proportion p
Standard deviation	$\sqrt{\frac{p\,(1-p)}{n}}$
Test Statistic	Sample Proportion $\hat{p}$

Tests of significance (Hypothesis test)

3. find and explain p-value

The p-value represents the probability of getting our test statistic or any test statistic more extreme **if the null hypothesis is true.**

For a one-sided "greater than" alternative hypothesis, the "more extreme" part refers to test statistic values larger than the test statistic given.

For a one-sided "less than" alternative hypothesis, the "more extreme" part of the interpretation refers to test statistic values smaller than the test statistic given.

4. decision and conclusion. Often $\alpha = 0.05$, unless otherwise specified

p-value $< \alpha$ => Reject H_0, Strong evidence for H_A

p-value $>= \alpha$=> Do not Reject H_0, insufficient evidence for H_A