STAT 103
Statistical Thinking

Exam 3 Review

Professor Mengshi Zhou

Chapter 14: Scatterplot

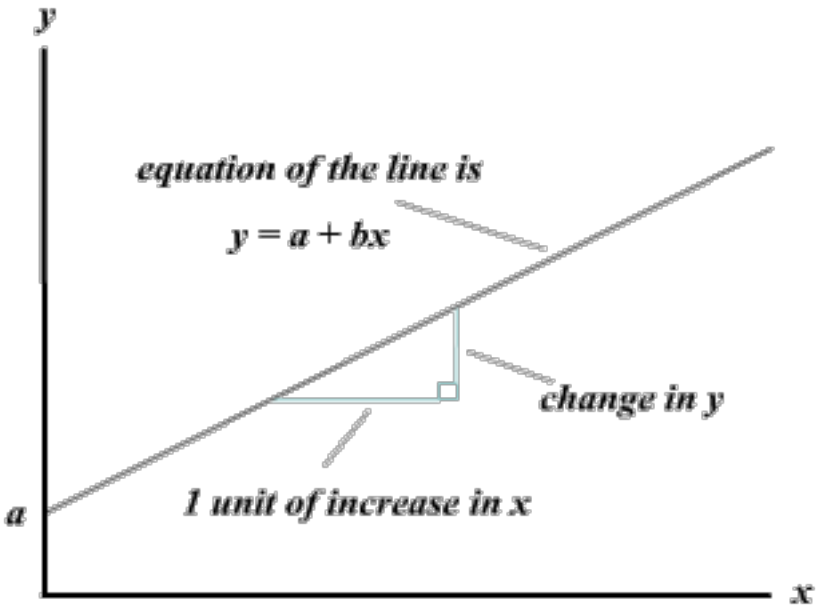
- How to read the scatter plot
- If there is a clear **direction**, is it positive (the scatterplot slopes upward from left to right) or negative (the plot slopes downward)?
- Is the **Form** straight (linear) or curved (non-linear)?
- **Strength:** How closely do the points follow the form (shape)
  - 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

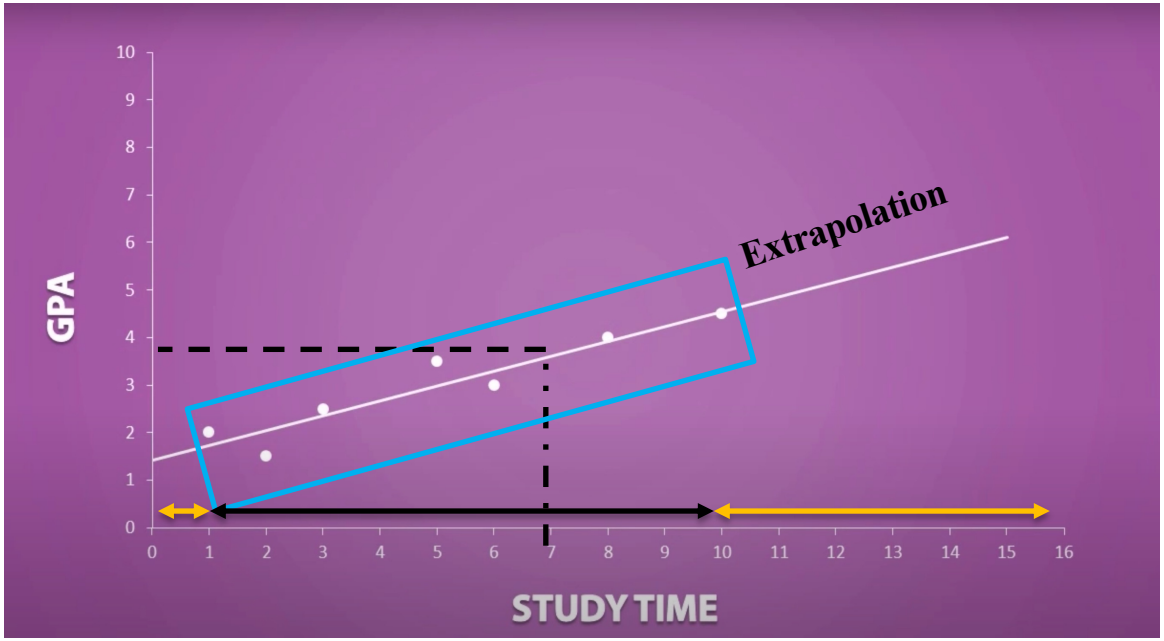


Regression Line Issues

- 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

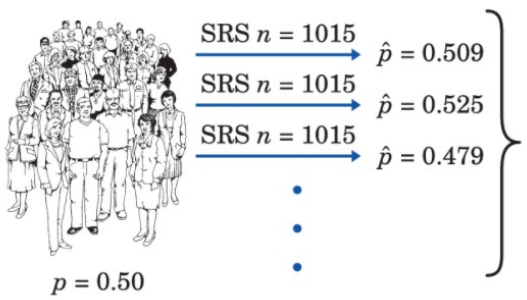
•Repeated Random Samples

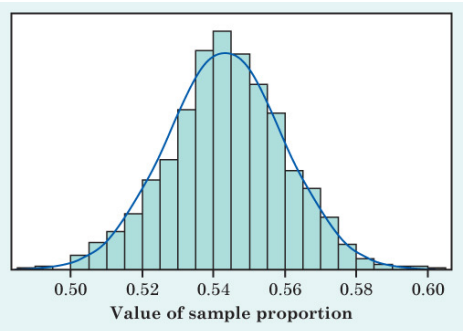
•In the long run, this distribution of statistics is well described by the normal curve.

•The mean of the normal curve is the value parameter

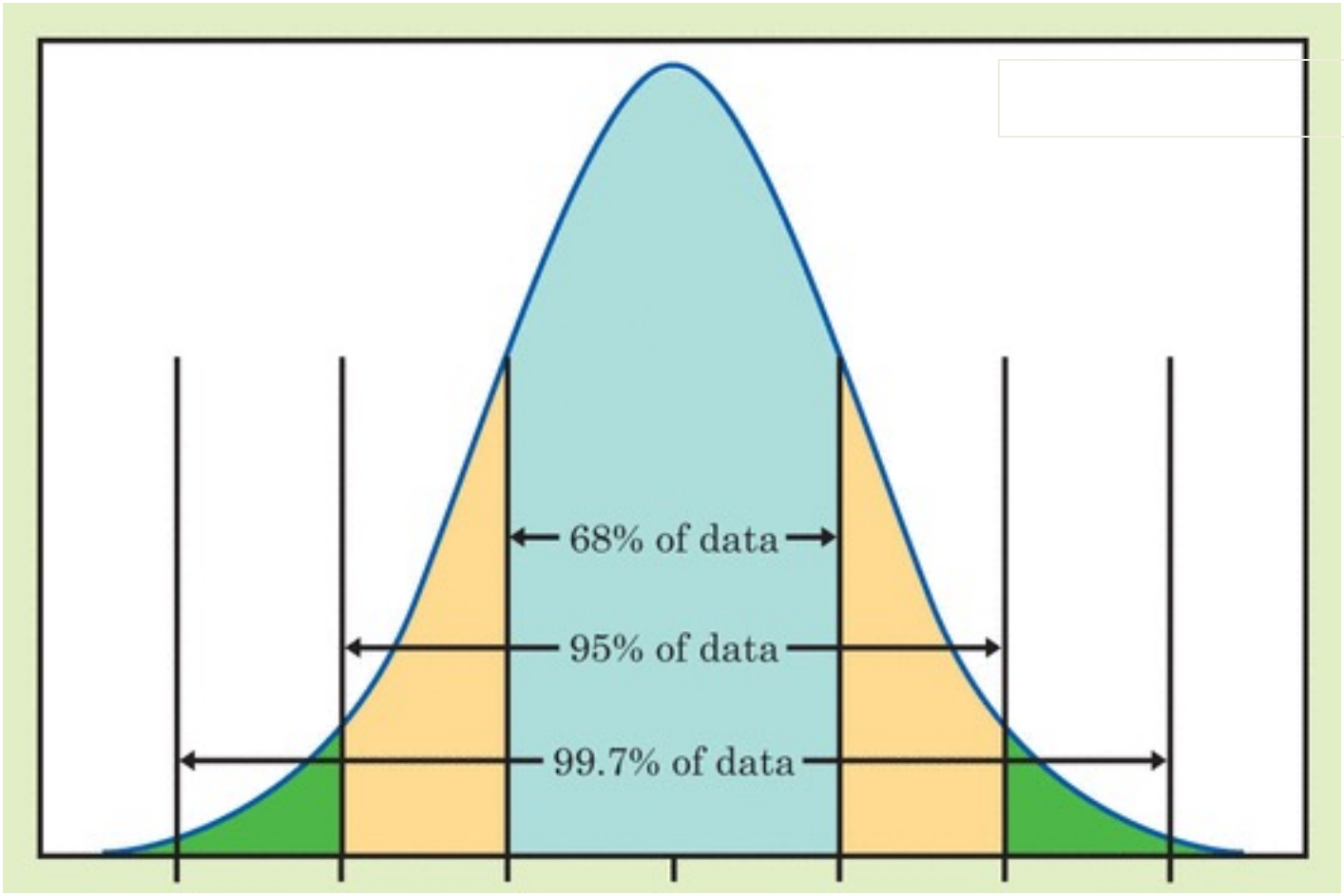
•The total probability is 1 because the total area under the curve is 1.

•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%