Phase 2 Expectations:

Statistics

Agenda

1. Phase 2 Overview

2. Week 1: SQL & Probability

3. Week 2: Hypothesis Testing & Simple Linear Regression
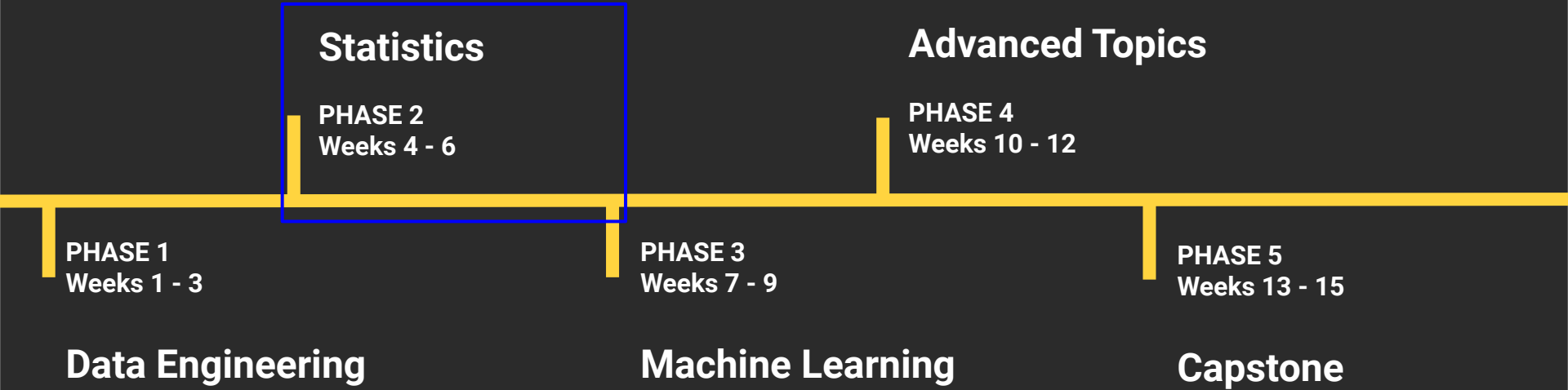
4. Gating & Assessments

5. Questions/Feedback

Phase 2: Overview

Overall Phase Timeline

Statistics

PHASE 2
Weeks 4 - 6

Advanced Topics

PHASE 4
Weeks 10 - 12

PHASE 1
Weeks 1 - 3

PHASE 3
Weeks 7 - 9

PHASE 5
Weeks 13 - 15

Data Engineering

Machine Learning

Capstone

Josh Wills
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Reply   Retweet   Favorite   More

9:55 AM - 3 May 12

// FLATIRON SCHOOL

Statistics for Data Science

SQL

- Common data storage method
- Relational databases
- New way to access information
- Critical skill
- Integrate with python - pandas

Set Theory & Probability

- Unions, intersections, joins
- Uncertainty, confidence
- Predictive probability
- Conditional probability
- Distributions

Statistics → Data Science

- Hypothesis testings: Z-test, t-test, ANOVA, Chi-Squared
- Using distributions
- Beginning of data modeling
- Simple Linear regressions
- Quantifying relationships

Probability

Probability of Simple Events

$$P(A) = \frac{n}{N} = \frac{\#\,outcomes\ in\ A}{\#\,outcomes\ in\ Sample\ Space}$$

Probability

Multiplication Rule

Independent Events

$$P(X \cap Y) = P(X) \cdot P(Y)$$

Dependent Events

$$P(X \cap Y) = P(Y) \cdot P(X \mid Y)$$

Bayes' Theorem

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$

Probability of Compound Events

Independent Events

$$P(A \text{ and } B) = P(A) \times P(B)$$
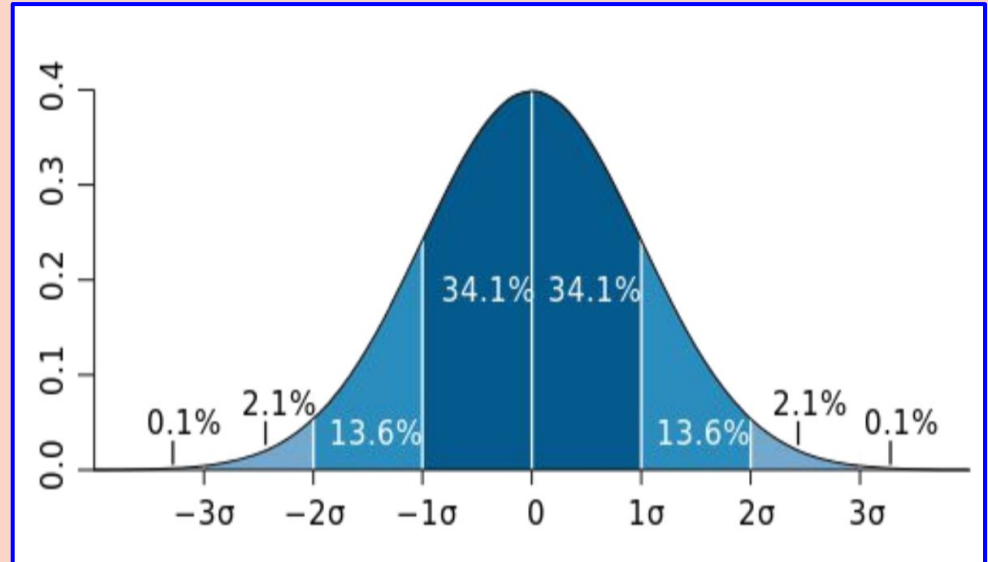
Dependent Events
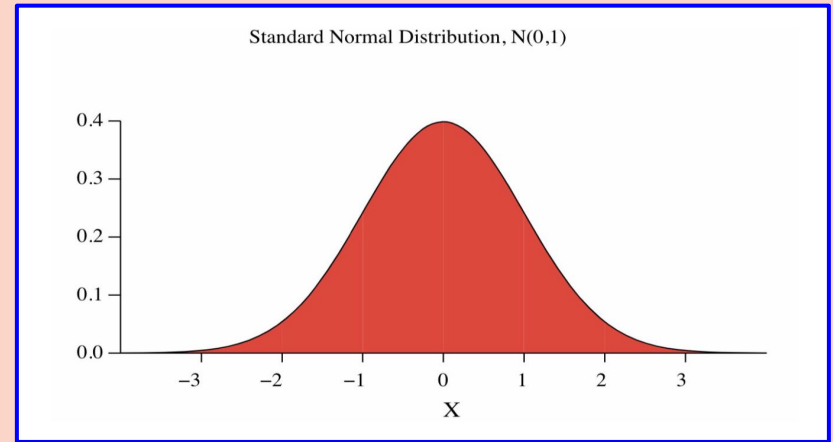
$$P(A \text{ and } B) = P(A) \times P(B \mid A)$$
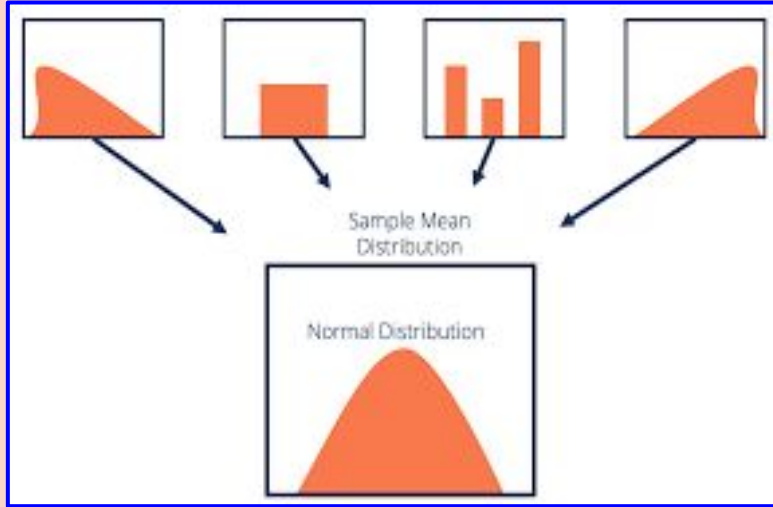
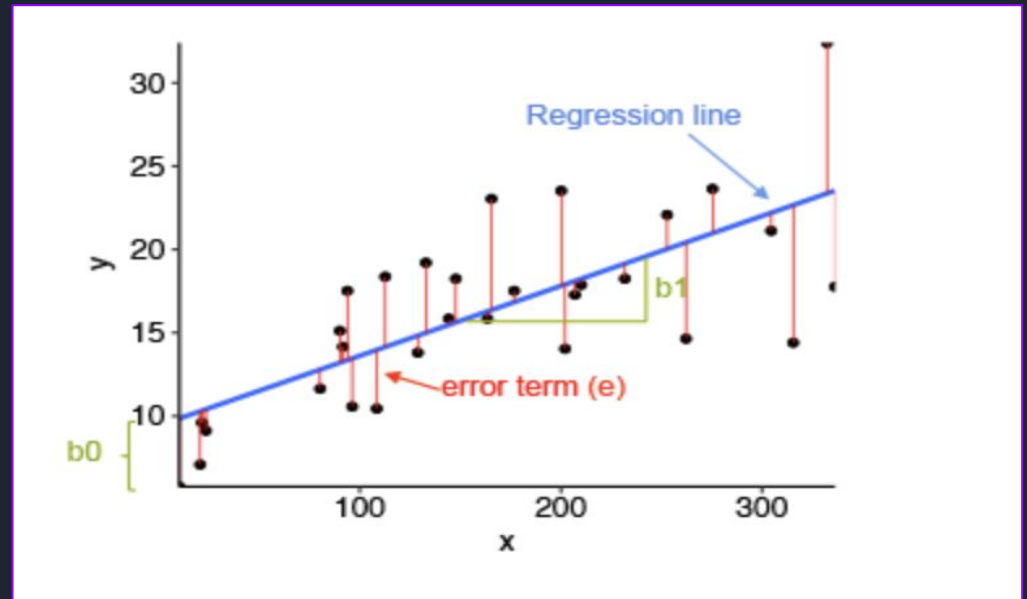Mutually Exclusive

$$P(A \text{ or } B) = P(A) + P(B)$$

Mutually Inclusive
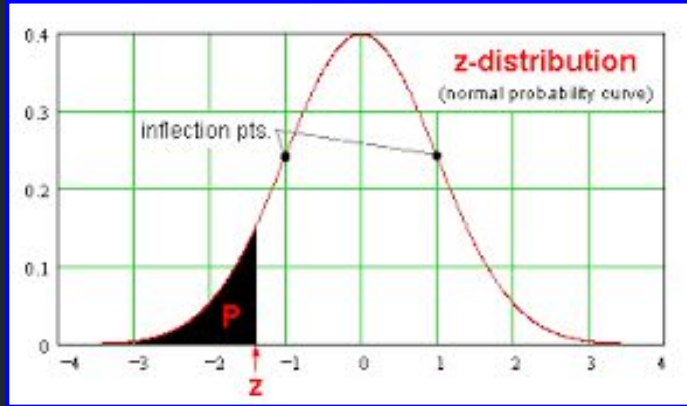
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Distributions

Sample Mean Distribution

Normal Distribution

Standard Normal Distribution, N(0,1)

34.1% 34.1%

0.1% 2.1% 13.6% 13.6% 2.1% 0.1%

-3σ -2σ -1σ 0 1σ 2σ 3σ

Hypothesis Testing

Measuring Student Progress

CP & CC	Blog Post	Project	1:1
7/19 - Wed - SQL Checkpoint 7/20 - Thurs - Probability Checkpoint 7/25 - Tues - Hypothesis Testing Checkpoint 7/28 - Friday - Phase 2 Code Challenge	- Due Tuesday of 2nd Week 7/25 - Topic: Describe a DS tool or library with tutorial - See canvas	- Group Project - Third week - Movie data - SQL database - Tiny bit of Lin Reg	Link to come later this week - Phase 1 review - Project review