Discrete Response Model
Lecture 1

datascience@berkeley

Introduction to Categorical Data, and the Bernoulli and Binomial Probability Models

What Is a Categorical (Qualitative) Variable?

- What is a categorical (qualitative) variable?
  - Patient survival: yes or no
  - Customer retention: churn or not
  - Produce color choice: blue, green, yellow, …
  - Self-reported health condition rating: 1,2,3,4,5
  - Customer satisfaction: Satisfied, Neutral, Unsatisfied
  - Highest attained education level : HS, BS, MS, PhD (ordinal properties)
  - Annual income: <15,000, 15,000-<25,000, 25,000-<40,000, ≥40,000 (ordinal properties)
- The first three examples do not have a natural ordering, and the last four examples have a natural ordering. We call them ordinal variables.

- We will focus on binary response in Lecture 1 and 2.

Binary Response Variable Observed From a Homogeneous Population

Goal: Estimate the overall probability of observing one of two possible outcomes for this random variable.

- This is often equated with the "probability of success" for an individual item in the population.
- Equivalently, this is the overall prevalence of successes in the population because each item has the same probability of success.

Bernoulli and Binomial Probability Distributions

- Suppose Y = 1 is a success where the probability of a success is P(Y = 1) = π, Y = 0 is a failure.

- Goal: Estimate π.

Bernoulli probability mass function:

P(Y = y) = $\pi^y(1-\pi)^{1-y}$ for y = 0 or 1

Notice that P(Y = 1) = π and P(Y = 0) = 1 − π

Often, you observe multiple success/failure observations. Let Y_1, …, Y_n denote random variables for these observations. If the random variables are independent and have the same probability of success π, then we can use a binomial PMF for $W = \sum_{i=1}^{n} Y_i$.

Binomial Probability Distributions

$$P(W = w) = \frac{n!}{w!(n-w)!}\pi^{w}(1-\pi)^{n-w}$$

for w = 0, 1, …, n

Notes:

- $\frac{n!}{w!(n-w)!} = \binom{n}{w}$ = n choose w
- W is a random variable denoting the number of "successes" out of n trials
- W has a fixed number of possibilities – 0, 1, …, n
- n is a fixed constant
- π is a parameter denoting the probability of a "success" with values between 0 and 1.

Required Conditions When Applying Binomial Probability Model

1. There are n identical trials.
2. Each trial has two possible outcomes, typically referred to as a success or failure.
3. The trials are independent of each other.
4. The probability of success, denoted by π, remains constant for each trial. The probability of a failure is $1-\pi$.
5. The random variable, W, represents the number of successes.

We will use two running (toy) examples to illustrate the concepts and techniques in this lecture.

As a reminder, please read the required reading before attending the live sessions.

An Example: Field Goal Kicking

Suppose a field goal kicker attempts five field goals during a game and each field goal has the same probability of being successful (the kick is made). Also, assume each field goal is attempted under similar conditions; i.e., distance, weather, surface,….

1. n identical trials: n = 5 field goals attempted under identical conditions.

2. Two possible outcomes of a trial: Each field goal can be made (success) or missed (failure).

3. The trials are independent of each other: The result of one field goal does not affect the result of another field goal.

An Example: Field Goal Kicking (cont.)

4. The probability of success, denoted by π, remains constant for each trial. The probability of a failure is $1-\pi$: Suppose the probability a field goal is good is 0.6; i.e., $P(\text{success}) = \pi = 0.6$.

5. The random variable, W, represents the number of successes. Let W = number of field goals that are good. Thus, W can be 0, 1, 2, 3, 4, or 5. Because these five items are satisfied, the binomial probability mass function can be used, and W is called a binomial random variable.

Mean and Variance of Binomial Probability Distributions

$$E(W) = n\pi$$

$$Var(W) = n\pi(1-\pi)$$

Berkeley SCHOOL OF INFORMATION