Introduction to Categorical Data Analysis

Grinnell College

January 21, 2025

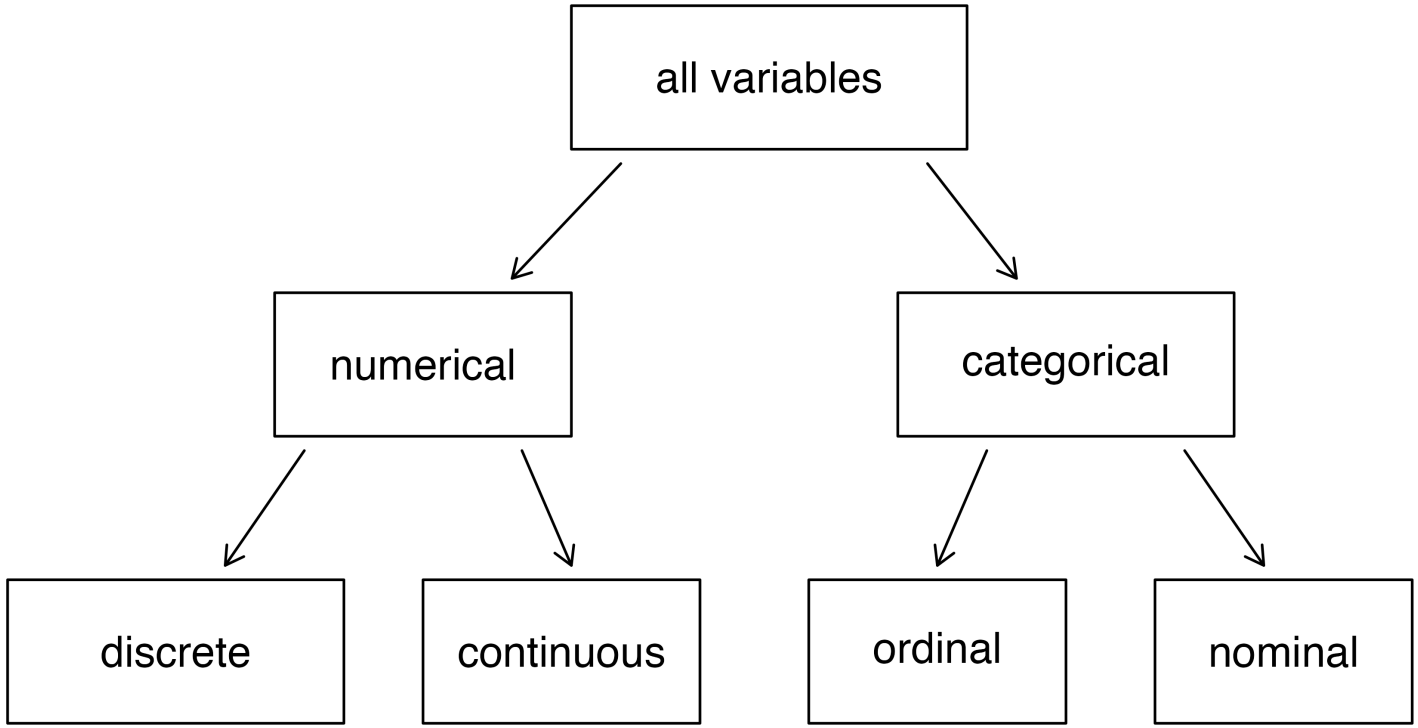
Goals today

- What is categorical data analysis?
- Types of data
- Notation for the class
- Distribution functions

"It is unnecessary and often inappropriate to use methods for continuous data with categorical responses"

- Inference on discrete data
- Categorical measures of association
- Tables, proportions, and odds ratios
- Generalized linear models
- Logistic regression

Taxonomy of Data



Categorical data is data that has as its measurement scale a *set of categories*

Categorical data itself has a few distinctions:

- Binary data, categories with only two outcomes (i.e., yes/no)
- Nominal data, a generalization of binary data in which there are multiple *unordered* categories (i.e., car, truck, SUV)
- Ordinal data which contains distinct categories with a clear hierarchical scale (i.e., high school, some college, college, graduate)

Response vs Explanatory variables

We often further classify variables according to the role they play in our statistical analyses:

The **response variable**, also called the *dependent variable*, is often associated with the goal of a study or analysis. In other words, this is typically the variable we are attempting to learn about. We call it *dependent* as its value often depends on the explanatory variables

The **explanatory variable**, or *independent variable*, typically represents characteristics of the entity we are studying and is used to predict or inform the values of the response variable

Notation

This class will use Y and X to refer to dependent and independent variables, respectively

A distinction will be made between Y, referring to a variable as a random, unrealized quantity, and y, a specific realization of a random variable

For example, Y may represent the outcome of a yet to be performed coin toss ($Y = \{0, 1\}$), whereas $y = 1$ refers to a specific instance in which the coin was tossed and landed on heads

Notation

Regarding random variables, we are often interested in discussing their **expected value** (the mean) and their **standard deviations**. These will be denoted as

$$E(Y) = \mu, \qquad \sigma(Y) = \sigma$$

We will use π to denote any probabilitiy

For example, then, a random binomial Y with n trials and probability of success π will have an expected value of

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

Random Variables and Distributions

A random variable, depending on its type, can be expected to take one of any number of values once it is *realized*

A description of *what values* a random variable will take, along with *how frequently they occur*, is referred to as a **distribution**

Particularly with categorical variables, the distribution that a variable takes will be implied by the process generating the outcomes

Independent and identically distributed random variables

When discussing a collection of objects, outcomes, or trials from a distribution, we often refer to them as being **independent** and **identically distributed**

That a collection of trials are *independent* means that the outcome of one trial has no bearing on the outcome of any others

A collection of trials are *identically distributed* if the data-generating process is the same for each of the observations

Together, we often refer to these as being *IID*

Probability Mass Functions

Generally speaking, a **parameter** refers to any type of numerical summary of a distribution (i.e., the mean, the mode)

A subset of these, known as **distributional parameters** are parameters such that, if known, communicate everything that can be known about a distribution

Probability Mass Functions

Mathematically, a distribution is described by a probability function called a **probability density function** (PDF) for continuous random variables and a **probability mass function** (PMF) for discrete random variables.

For example, a **binomial** distribution has distributional parameters n and π (written Bin(n, π)):

$$P(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

valid for all values $y = 0, 1, 2, \ldots, n$

Probability Mass Functions

For example, suppose we have a binomial distribution with $n = 8$ trials and probability of success $\pi = 0.5$

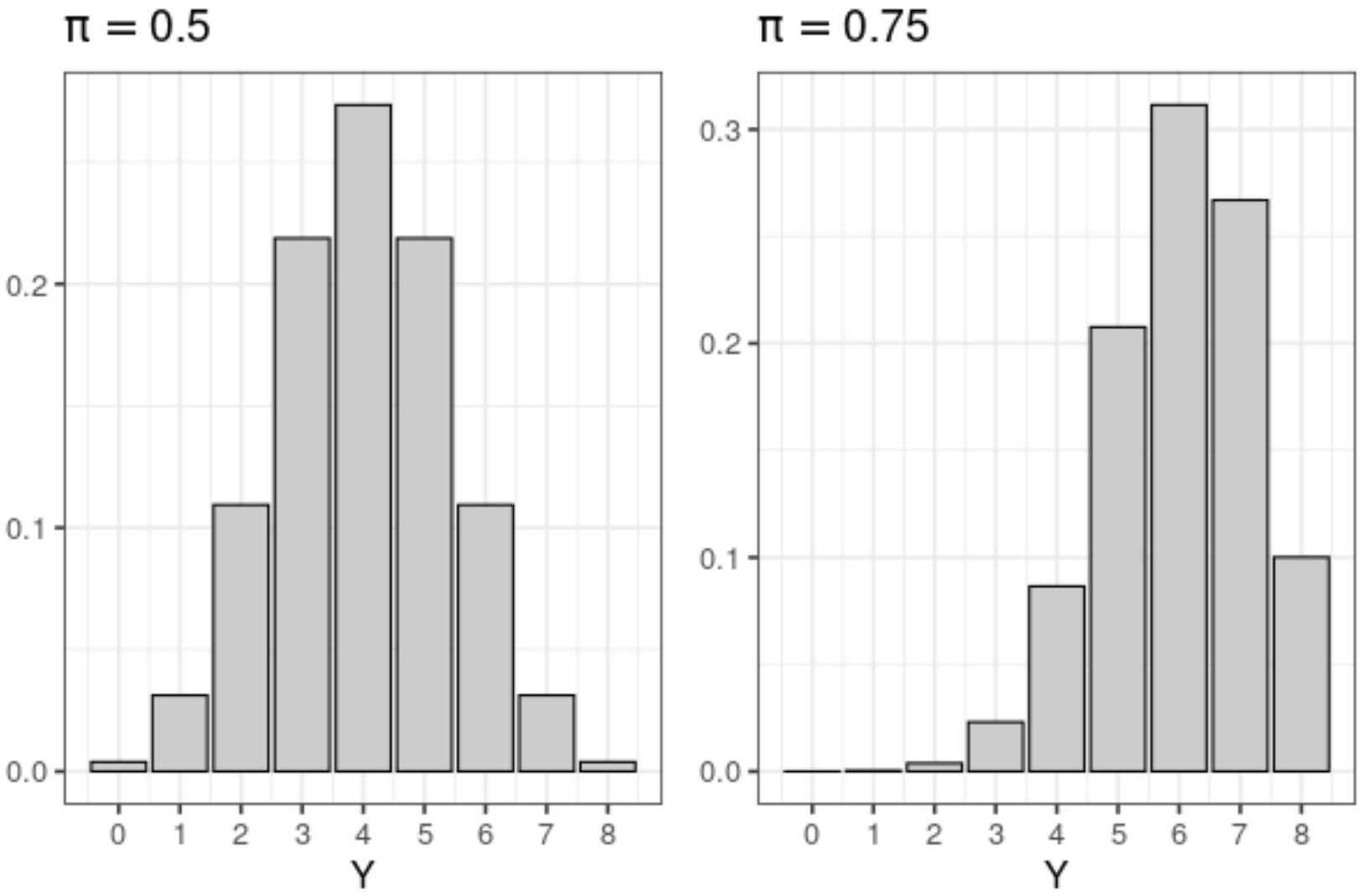
The probability that we observe 6 successes is

$$P(Y = 6) = \binom{8}{6}(0.5)^6(1 - 0.5)^2$$

$$= \frac{8!}{6!(8-6)!}(0.5)^8$$

$$= 0.109$$

Probability Mass Functions



Recap

Words words words