STATS 111/202

Lecture 1: Categorical variables and common distributions

1/6/2025

Ana Maria Kenney

UC Irvine Department of Statistics

STAT 110/201 vs 111/202

STAT 110/201

- Study how independent/explanatory variables X explain the variance in responses Y
- X is a quantitative variable (continuous) or qualitative (discrete/categorical)
- Y (strictly) continuous

STAT 110/201 vs 111/202

STAT 110/201

- Study how independent/explanatory variables X explain the variance in responses Y
- X is a quantitative variable (continuous) or qualitative (discrete/categorical)
- Y (strictly) continuous

Linear regression

- Assumed Y is Normally distributed
- $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
- Assume errors, ε independent and Normal(0,σ_ε^2)
- Implies also Normal with mean $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ and variance σ_ε^2

STAT 110/201 vs 111/202

STAT 110/201

- Study how independent/explanatory variables X explain the variance in responses Y
- X is a quantitative variable (continuous) or qualitative (discrete/categorical)
- Y (strictly) continuous

Linear regression

- Assumed Y is Normally distributed
- $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
- Assume errors, ε independent and Normal(0,σ_ε^2)
- Implies also Normal with mean $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ and variance σ_ε^2

Goal: estimate β's and σ_ε^2

STAT 110/201 vs 111/202

STAT 110/201

- Study how independent/explanatory variables X explain the variance in responses Y
- X is a quantitative variable (continuous) or qualitative (discrete/categorical)
- **Y (strictly) continuous**

Linear regression

- Assumed Y is Normally distributed
- $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
- Assume errors, ε independent and Normal($0,\sigma_\varepsilon^2$)
- Implies also Normal with mean $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ and variance σ_ε^2

How do we generalize to different types of Y?
Specifically, if Y is categorical with 2 or more categories

Types of categorical variables

1.1

Types of categorical variables

Categorical variables: support has a discrete set of values/categories

- Blood type: A, B, AB, O
- Eye color: brown, green blue
- Letter grade: A, B, C, D, F

Types of categorical variables

Nominal categorical variables: outcomes have no inherent ordering
Ordinal categorical variables: outcomes have an inherent ordering

Label these as nominal, ordinal, or continuous

- Education level (high school, some college, college, graduate school)
- Favorite building on campus
- Age
- IMDB ratings
- Number of children in a household

Problem: what is the distribution of Y?

STAT 110/201

- Assumed Y follows a **Normal distribution**
- Its support is on the real line – Y can be **any** number
- Specifically, Y is Normal with:
  - $E(Y) = \mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
  - $Var(Y) = \sigma_\varepsilon^2$

Common distributions

1.2, 1.3

Binomial distribution

If Y follows a binomial distribution:

- Two important parameters: n and p
  - n: Number of independent trials with binary outcome
  - p: Probability of success on a given trial
- A binomial random variable is the number of successful trials
- These independent trials are often called **Bernoulli trials**
- The probability mass function (pmf) is
  - $f(y) = P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$
- The expectation and variance is
  - $E(Y) = np$
  - $Var(Y) = np(1-p)$

Called the binomial coefficient: number of ways to have y successes

Binomial distribution

If Y follows a binomial distribution:

- Two important parameters: n and p
  - n: Number of independent trials with binary outcome
  - p: Probability of success on a given trial
- A binomial random variable is the number of successful trials
- These independent trials are often called **Bernoulli trials**
- The probability mass function (pmf) is
  - $f(y) = P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$
- The expectation and variance is
  - $E(Y) = np$
  - $Var(Y) = np(1-p)$

Example: Let Y be the number of correct answers in a 10 question, multiple-choice quiz (5 options per question)

Called the binomial coefficient: number of ways to have y successes

Binomial distribution

If Y follows a binomial distribution:

- Two important parameters: n and p
  - n: Number of independent trials with binary outcome
  - p: Probability of success on a given trial
- A binomial random variable is the number of successful trials
- These independent trials are often called **Bernoulli trials**
- The probability mass function (pmf) is
  - $f(y) = P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$
- The expectation and variance is
  - $E(Y) = np$
  - $Var(Y) = np(1-p)$

Want to model this

Example: Let Y be the number of correct answers in a 10 question, multiple-choice quiz (5 options per question)

Bernoulli distribution

- When $n = 1$, Y follows a Bernoulli distribution.
- Single trial, outcome is binary (success or not)
  - Success? Yes → $Y = 1$
  - Success? No → $Y = 0$
- One important parameters: p
  - p: Probability of success
  - $P(Y = 1) = p$
  - $P(Y = 0) = 1 - p$
- The expectation and variance is
  - $E(Y) = p$
  - $Var(Y) = p(1 - p)$

Can use to model yes/no responses

Multinomial distribution

What if each trial can result in more than two outcomes?

- Multinomial returns the number of times each category occurred out of the n random trials
- Let $k > 2$ be the number of outcomes/categories for a given trial
- Let X_i be the number of times the i-th category occurred out of n trials ($i = 1,2, \dots, k$)
  - $\sum_{i=1}^{k} X_k = n$
- Let $p_1, p_2, \dots, p_k$ be the probabilities of each category
  - $\sum_{i=1}^{k} p_i = 1$
- Then the probability mass function is:
  - $f(x_1, x_2, \dots, x_k) = P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$
- $E(X_i) = np_i$
- $Var(X_i) = np_i(1 - p_i)$
- $Cov(X_i, X_j) = -np_i p_j$

Multinomial distribution

What if each trial can result in more than two outcomes?

- Multinomial returns the number of times each category occurred out of the n random trials
- Let $k > 2$ be the number of outcomes/categories for a given trial
- Let X_i be the number of times the i-th category occurred out of n trials ($i = 1,2,\ldots,k$)
  - $\sum_{i=1}^{k} X_k = n$
- Let $p_1, p_2, \ldots, p_k$ be the probabilities of each category
  - $\sum_{i=1}^{k} p_i = 1$
- Then the probability mass function is:
  - $f(x_1, x_2, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$
- $E(X_i) = np_i$
- $Var(X_i) = np_i(1 - p_i)$
- $Cov(X_i, X_j) = -np_ip_j$

Want to estimate these probabilities and (eventually) study how they change

Multinomial distribution

What if each trial can result in more than two outcomes?

- Multinomial returns the number of times each category occurred out of the n random trials
- Let $k > 2$ be the number of outcomes/categories for a given trial
- Let X_i be the number of times the i-th category occurred out of n trials ($i = 1,2,\dots,k$)
  - $\sum_{i=1}^{k} X_k = n$
- Let $p_1, p_2, \dots, p_k$ be the probabilities of each category
  - $\sum_{i=1}^{k} p_i = 1$
- Then the probability mass function is:
  - $f(x_1, x_2, \dots, x_k) = P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$
- $E(X_i) = np_i$
- $Var(X_i) = np_i(1 - p_i)$
- $Cov(X_i, X_j) = -np_i p_j$

Binomial and Bernoulli distributions are special cases of the multinomial!

Multinomial distribution

What if each trial can result in more than two outcomes?

- Multinomial returns the number of times each category occurred out of the n random trials
- Let $k > 2$ be the number of outcomes/categories for a given trial
- Let X_i be the number of times the i-th category occurred out of n trials ($i = 1,2,\ldots,k$)
  - $\sum_{i=1}^{k} X_k = n$
- Let $p_1, p_2, \ldots, p_k$ be the probabilities of each category
  - $\sum_{i=1}^{k} p_i = 1$
- Then the probability mass function is:
  - $f(x_1, x_2, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$
- $E(X_i) = np_i$
- $Var(X_i) = np_i(1 - p_i)$
- $Cov(X_i, X_j) = -np_i p_j$ Why negative?

Multinomial distribution

What if each trial can result in more than two outcomes?

- Multinomial returns the number of times each category occurred out of the n random trials
- Let $k > 2$ be the number of outcomes/categories for a given trial
- Let X_i be the number of times the i-th category occurred out of n trials $(i = 1,2,\ldots,k)$
  - $\sum_{i=1}^{k} X_k = n$
- Let $p_1, p_2, \ldots, p_k$ be the probabilities of each category
  - $\sum_{i=1}^{k} p_i = 1$
- Then the probability mass function is:
  - $f(x_1, x_2, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$
- $E(X_i) = np_i$
- $Var(X_i) = np_i(1 - p_i)$
- $Cov(X_i, X_j) = -np_i p_j$

Hardest (most interesting case), will swing back later in the quarter…

Hypergeometric distribution

Say we have N objects/trials with K successes. What if we randomly select n of them?

- Three important parameters: N, K, and n
- Let X be the number of successes from the n draws
  - X can take on values (i.e., has support) $\{\max(0, n+K-N), \ldots, \min(n, K)\}$
- The probability mass function is:
  - $f(x) = P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$

Hypergeometric distribution

Say we have N objects/trials with K successes. What if we randomly select n of them?

- Three important parameters: N, K, and n
- Let X be the number of successes from the n draws
  - X can take on values (i.e., has support) $\{\max(0, n + K - N), \ldots, \min(n, K)\}$
- The probability mass function is:
  - $f(x) = P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$

Example: Say everyone in the class takes a flu test and 5 of them are defective. If we randomly select 4 without replacement, then take X = the number of defective tests.

Poisson distribution

What if we want to study the number of times something happens

- Examples: count data, events occurring randomly over a time period
- Just have one important parameter $\lambda > 0$
  - This is the expected rate of these events occurring
- Take X following a Poisson distribution with parameter λ
  - The support of X is $\{0,1,2, \ldots, \infty\}$
- Then the probability mass function is:
  - $f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$
- $E(X) = \lambda$
- $Var(X) = \lambda$

Poisson distribution

What if we want to study the number of times something happens

- Examples: count data, events occurring randomly over a time period
- Just have one important parameter $\lambda > 0$
  - This is the expected rate of these events occurring
- Take X following a Poisson distribution with parameter λ
  - The support of X is $\{0,1,2,\ldots,\infty\}$
- Then the probability mass function is:
  - $f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$
- $E(X) = \lambda$
- $Var(X) = \lambda$

Example: Let X be the number of angry tweets a statistics professor sends out the first 24 hours after a bad analysis gets in Nature. Previous occurrences show they tweet an average of 53 every 24 hours.

Poisson distribution: cool properties

Infinitely divisible distribution

- In general, if a random variable (r.v.) is infinitely divisible, then we can write it as the sum of n many independent random variables (n can be infinitely large)
- For a Poisson with parameter λ, it can be written as the sum of n independent Poisson random variables with parameter $\frac{\lambda}{n}$
  - Also implies the sum of independent Poisson r.v. with different parameters λ_i is also a Poisson with parameter $\sum_i \lambda_i$
- Also means we can shift to different time intervals!

Poisson distribution: cool properties

Infinitely divisible distribution

- In general, if a random variable (r.v.) is infinitely divisible, then we can write it as the sum of n many independent random variables (n can be infinitely large)
- For a Poisson with parameter λ, it can be written as the sum of n independent Poisson random variables with parameter $\frac{\lambda}{n}$
  - Also implies the sum of independent Poisson r.v. with different parameters λ_i is also a Poisson with parameter $\sum_i \lambda_i$
- Also means we can shift to different time intervals!

May make more sense to model with different rates. For instance, the average number of tweets sent at 3pm will differ from 3am

Poisson distribution: cool properties

Infinitely divisible distribution

- In general, if a random variable (r.v.) is infinitely divisible, then we can write it as the sum of n many independent random variables (n can be infinitely large)
- For a Poisson with parameter λ, it can be written as the sum of n independent Poisson random variables with parameter $\frac{\lambda}{n}$
  - Also implies the sum of independent Poisson r.v. with different parameters λ_i is also a Poisson with parameter $\sum_i \lambda_i$
- Also means we can shift to different time intervals!

Example: Let X be the number of angry tweets a statistics professor sends out the first 24 hours after a bad analysis gets in Nature. Previous occurrences show they tweet an average of 53 every 24 hours.

Announcements

- Canvas will be unlocked later today
- Homework 1 will be posted tonight (tomorrow at the latest) and **due next Friday January 17th at 11:59PM PT**