Lecture 1: Course Introduction, Basics of Probability and Statistics, & A Little Bit about Contingency Tables

STATS305B: Applied Statistics II

Scott Linderman

January 6, 2025

Introductions

- **Instructor:** Scott Linderman (Asst. Prof., Statistics and Wu Tsai Neuro. Inst.)
- **TA:** Amber Hu (PhD student, Statistics)
- **TA:** Michael Salerno (PhD student, Statistics)

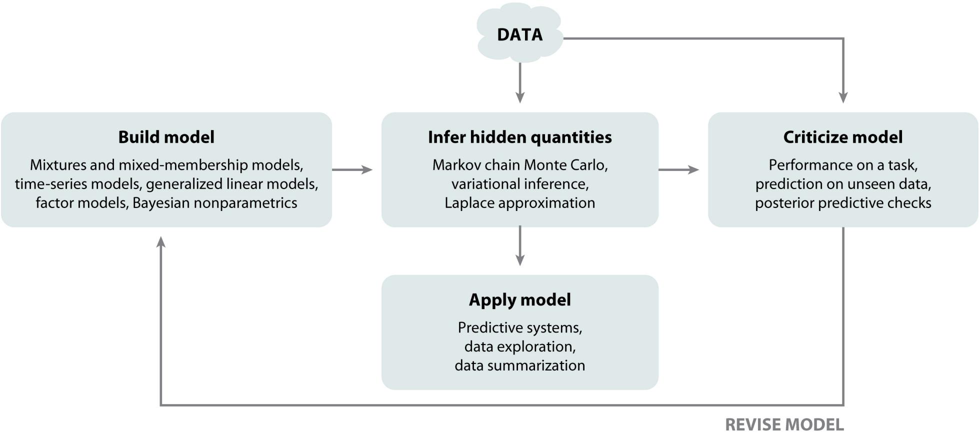
What is this course about?

Probabilistic modeling and inference with discrete data.

To what end? We want to:

- **Predict:** given features, estimate labels or outputs
- **Simulate:** given partial observations, generate the rest
- **Summarize:** given high dimensional data, find low-dimensional factors of variation
- **Decide:** given past actions/outcomes, which choice is best?
- **Understand:** what generative mechanisms gave rise to this data?

Box’s Loop



DATA

Build model

Mixtures and mixed-membership models, time-series models, generalized linear models, factor models, Bayesian nonparametrics

Infer hidden quantities

Markov chain Monte Carlo, variational inference, Laplace approximation

Criticize model

Performance on a task, prediction on unseen data, posterior predictive checks

Apply model

Predictive systems, data exploration, data summarization

REVISE MODEL

Blei DM. 2014.
Annu. Rev. Stat. Appl. 1:203–32

Course Outline

- Weeks 1-3: Classics: Exponential family distributions and GLMs
- Weeks 4-5: Bayesian Inference algorithms: MCMC and variational inference
- Weeks 6-7: Latent variable models: mixture models, HMMs, etc.
- Weeks 8-9: Deep generative models: VAEs, Transformers, Deep SSMs, Denoising diffusion models
- Week 10: Bonus: Point processes, survival analysis, etc.

Learning Objectives

- Understand the mathematical underpinnings of classical and modern models for discrete data.
- Develop expertise in an array of algorithms for parameter estimation and inference in these models.
- Be able to code these models and algorithms from scratch in Python.

Assignments

- Weeks 1-3: Classics: Exponential family distributions and GLMs

  Predict outcomes of college football games with a Bradley-Terry model.

- Weeks 4-5: Bayesian Inference algorithms: MCMC and variational inference

  Election forecasting with a Bayesian GLM.

- Weeks 6-7: Latent variable models: mixture models, HMMs, etc.

  Changepoint detection in time series data.

- Weeks 8-9: Deep generative models: VAEs, Transformers, Deep SSMs, Denoising diffusion models

  Build a small LLM.

- Week 10: Bonus: Point processes, survival analysis, etc.

Logistics

Please see the course website for the syllabus, schedule, lecture notes, grading policy, etc.

https://slinderman.github.io/stats305b/

We will use Ed for communications, questions, etc., and Gradescope for assignments.

Interactive Example: College Football Playoffs

Let's start with a hands-on example. The College Football Playoffs are underway, and the Super Bowl is coming up in a few weeks! If you go to a watch party, you might like to play the following game with your friends.

Before the football game starts, create a 10x10 board with the rows and columns numbered 0 through 9. Each cell represents a possible final score of the home and away team, mod 10. You and your friends select cells in round robin order until all 100 cells are taken. Whoever has the cell with the final score (mod 10) wins!

First, let's review some basics about football...

Step 1: Collect Data

Let’s play together, using the upcoming Cotton Bowl between Ohio State and Texas as our example. Fill out this poll to enter your guess.

`https://tinyurl.com/stats305lec01`

Outline

- Basic Probability Distributions
- Maximum Likelihood Estimation
- Contingency Tables and Independence

Bernoulli Distribution

Toss a (biased) coin where the probability of heads is $p \in [0,1]$. Let $X = 1$ denote the event that a coin flip comes up heads and $X = 0$ it comes up tails. The random variable X follows a Bernoulli distribution,

$$X \sim \mathrm{Bern}(p)$$

We denote its **probability mass function (pmf)** by,

$$\mathrm{Bern}(x; p) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

or more succinctly

$$\mathrm{Bern}(x; p) = p^x (1 - p)^{1-x}.$$

The Bernoulli distribution's mean is $\mathbb{E}[X] = p$ and its variance is $\mathrm{Var}[X] = p(1 - p)$.

Binomial Distribution

Now toss the same biased coin n times independently and let

$$X_i \overset{\text{iid}}{\sim} \mathrm{Bern}(p) \qquad \text{for } i = 1, \dots, n$$

denote the outcomes of each trial.

The number of heads, $X = \sum_{i=1}^{n} X_i$, is a random variable taking values $X \in \{0, \dots, n\}$. It follows a binomial distribution,

$$X \sim \mathrm{Bin}(n, p)$$

with pmf

$$\mathrm{Bin}(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

Its mean and variance are $\mathbb{E}[X] = np$ and $\mathrm{Var}[X] = np(1 - p)$, respectively.

Poisson Distribution

Now let $n \rightarrow \infty$ and $p \rightarrow 0$ while the product $np = \lambda$ stays constant. In that limit, the binomial distribution converges to the Poisson distribution over non-negative integers $X \in \mathbb{N}$,

$$X \sim \mathrm{Po}(\lambda).$$

Its pmf is,

$$\mathrm{Po}(x; \lambda) = \frac{1}{x!} e^{-\lambda} \lambda^{x}.$$

The mean and variance are both λ. The fact that the mean equals the variance is a defining property of the Poisson distribution, but it's not always an appropriate modeling assumption.

Categorical Distribution

Instead of a biased coin, consider a biased *die* with K faces and corresponding probabilites $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \in \Delta_{K-1}$, where Δ_{K-1} denotes the $(K-1)$-dimensional simplex,

$$\Delta_{K-1} = \left\{ \boldsymbol{\pi} \in \mathbb{R}_+^K : \sum_k \pi_k = 1 \right\}$$

The outcome $X \in \{1, \ldots, K\}$ follows a categorical distribution, $X \sim \mathrm{Cat}(\boldsymbol{\pi})$, with pmf

$$\mathrm{Cat}(x; \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{\mathbb{I}[x=k]}$$

where $\mathbb{I}[y]$ is an indicator function that returns 1 if y is true and 0 otherwise.

The categorical distribution is a natural generalization of the Bernoulli distribution to random variables that can fall into more than two categories.

Categorical Distribution

Alternatively, we can represent X as a **one-hot vector**, in which case $X \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K\}$ where $\boldsymbol{e}_k = (0, \ldots, 1, \ldots, 0)^\top$ is a one-hot vector with a 1 in the k-th position. Then, the pmf is,

$$\mathrm{Cat}(\boldsymbol{x}; \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{x_k}$$

Multinomial Distribution

From this representation, it is straightforward to generalize to n independent rolls of the die, just like in the binomial distribution. Let $Z_i \overset{\text{iid}}{\sim} \mathrm{Cat}(\boldsymbol{\pi})$ for $i = 1, \ldots, n$ denote the outcomes of each roll, and let $X = \sum_{i=1}^{n} Z_i$ denote the total number of times the die came up on each of the K faces. Note that $X \in \mathbb{N}^K$ is a *vector-valued random variable*. Then, X follows a multinomial distribution,

$$X \sim \mathrm{Mult}(n, \boldsymbol{\pi}),$$

with pmf,

$$\mathrm{Mult}(\boldsymbol{x}; n, \boldsymbol{\pi}) = \mathbb{I}[\boldsymbol{x} \in \mathcal{X}_n] \cdot \binom{n}{x_1, \ldots, x_K} \prod_{k=1}^{K} \pi_k^{x_k}$$

where $\mathcal{X}_n = \left\{\boldsymbol{x} \in \mathbb{N}^K : \sum_{k=1}^{K} x_k = n\right\}$ and $\binom{n}{x_1, \ldots, x_K} = \frac{n!}{x_1! \cdots x_K!}$ denotes the multinomial function.

Multinomial Distribution

The expected value of a multinomial random variable is $\mathbb{E}[X] = n\boldsymbol{\pi}$ and the $K \times K$ covariance matrix is,

$$\mathrm{Cov}(X) = n \begin{bmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \dots & -\pi_1\pi_K \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & \dots & -\pi_2\pi_K \\ \vdots & \vdots & \vdots & \vdots \\ -\pi_K\pi_1 & -\pi_K\pi_2 & \dots & \pi_K(1-\pi_K) \end{bmatrix}$$

with entries

$$[\mathrm{Cov}[X]]_{ij} = \begin{cases} n\pi_i(1-\pi_i) & \text{if } i = j \\ -n\pi_i\pi_j & \text{if } i \neq j \end{cases}$$

Poisson / Multinomial Connection

Suppose we have a collection of independent (but not identically distributed) Poisson random variables,

$$X_i \sim \mathrm{Po}(\lambda_i) \qquad \text{for } i = 1, \ldots, K.$$

Due to their independence, the sum $X_\bullet = \sum_{i=1}^K X_i$ is Poisson distributed as well,

$$X_\bullet \sim \mathrm{Po}\left(\sum_{i=1}^K \lambda_k\right)$$

(We'll use this $X_\bullet$ notation more when we talk about contingency tables.)

Conditioning on the sum renders the counts *dependent*. (They have to sum to a fixed value, so they can't be independent!) Specifically, given the sum, the counts follow a multinomial distribution,

$$(X_1, \ldots, X_K) \mid X_\bullet = n \sim \mathrm{Mult}(n, \pi)$$

Poisson / Multinomial Connection

where

$$\pi = \left(\frac{\lambda_1}{\lambda_\bullet}, \ldots, \frac{\lambda_K}{\lambda_\bullet}\right)$$

with $\lambda_\bullet = \sum_{i=1}^{K} \lambda_i$. In words, given the sum, the vector of counts is multinomially distributed with probabilities given by the normalized rates.

Outline

- Basic Probability Distributions
- **Maximum Likelihood Estimation**
- Contingency Tables and Independence

Maximum Likelihood Estimation

The distributions above are simple probability models with one or two parameters. How can we estimate those parameters from data? We'll focus on maximum likelihood estimation.

The log likelihood is the probability of the data viewed as a function of the model parameters $\boldsymbol{\theta}$. Given i.i.d. observations $\{x_i\}_{i=1}^n$, the log likelihood reduces to a sum,

$$\mathscr{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(x_i; \theta).$$

The maximum likelihood estimate (MLE) is a maximum of the log likelihood,

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\ \max \mathscr{L}(\boldsymbol{\theta})$$

(Assume for now that there is a single global maximum.)

Example: MLE for the Bernoulli distribution

Consider a Bernoulli distribution unknown parameter $\theta \in [0,1]$ for the probability of heads. Suppose we observe n independent coin flips

$$X_i \overset{\text{iid}}{\sim} \mathrm{Bern}(\theta) \qquad \text{for } i = 1, \ldots, n.$$

Observing $X_i = x_i$ for all i, the log likelihood is,

$$\begin{aligned} \mathscr{L}(\theta) &= \sum_{i=1}^{n} x_i \log \theta + (1 - x_i) \log(1 - \theta) \\ &= x \log \theta + (n - x) \log(1 - \theta) \end{aligned}$$

where $x = \sum_{i=1}^{n} x_i$ is the number of flips that came up heads.

Example: MLE for the Bernoulli distribution

Taking the derivative with respect to p,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}\mathscr{L}(\theta) &= \frac{x}{\theta} - \frac{n-x}{1-\theta} \\ &= \frac{x-n\theta}{\theta(1-\theta)}.\end{aligned}$$

Setting this to zero and solving for θ yields the MLE,

$$\hat{\theta}_{\mathsf{MLE}} = \frac{x}{n}.$$

Intuitively, the maximum likelihood estimate is the fraction of coin flips that came up heads. Note that we could have equivalently expressed this model as a single observation of $X \sim \mathrm{Bin}(n, \theta)$ and obtained the same result.

Asymptotic Normality of the MLE

If the data were truly generated by i.i.d. draws from a model with parameter $\boldsymbol{\theta}^\star$, then under certain conditions the MLE is asymptotically normal and achieves the Cramer-Rao lower bound,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathsf{MLE}} - \boldsymbol{\theta}^\star) \rightarrow \mathcal{N}(0, \mathcal{I}(\boldsymbol{\theta}^\star)^{-1})$$

in distribution, where $\mathcal{I}(\boldsymbol{\theta})$ is the **Fisher information matrix**. We obtain standard error estimates by taking the square root of the diagonal elements of the inverse Fisher information matrix and dividing by $\sqrt{n}$.

Fisher Information Matrix

The Fisher information matrix is the covariance of the **score function**, the partial derivative of the log likelihood with respect to the parameters. It’s easy to confuse yourself with poor notation, so let’s try to derive it precisely.

The log probability is a function that maps two arguments (a data point and a parameter vector) to a scalar, $\log p : \mathscr{X} \times \Theta \mapsto \mathbb{R}$. The score function is the partial derivative with respect to the parameter vector, which is itself a function, $\nabla_\theta \log p : \mathscr{X} \times \Theta \mapsto \Theta$.

Fisher Information Matrix

Now consider $\boldsymbol{\theta}$ fixed and treat $X \sim p(\cdot; \boldsymbol{\theta})$ as a random variable. The expected value of the score is zero,

$$\begin{aligned}
\mathbb{E}[\nabla_\theta \log p(X; \boldsymbol{\theta})] &= \int_{\mathcal{X}} p(x; \boldsymbol{\theta}) \nabla_\theta \log p(x; \boldsymbol{\theta})\,\mathrm{d}x \\
&= \int_{\mathcal{X}} p(x; \boldsymbol{\theta}) \frac{\nabla_\theta p(x; \boldsymbol{\theta})}{p(x; \boldsymbol{\theta})}\,\mathrm{d}x \\
&= \nabla_\theta \int_{\mathcal{X}} p(x; \boldsymbol{\theta})\,\mathrm{d}x \\
&= \nabla_\theta 1 \\
&= \mathbf{0}.
\end{aligned}$$

Fisher Information Matrix

The Fisher information matrix is the **covariance of the score function**, again treating $\boldsymbol{\theta}$ as fixed,

$$\begin{aligned}
\mathscr{I}(\boldsymbol{\theta}) &= \mathrm{Cov}[\nabla_\theta \log p(X;\boldsymbol{\theta})] \\
&= \mathbb{E}[\nabla_\theta \log p(X;\boldsymbol{\theta}) \nabla_\theta \log p(X;\boldsymbol{\theta})^\top] - \underbrace{\mathbb{E}[\nabla_\theta \log p(X;\boldsymbol{\theta})]}_{\mathbf{0}} \underbrace{\mathbb{E}[\nabla_\theta \log p(X;\boldsymbol{\theta})]^\top}_{\mathbf{0}^\top} \\
&= \mathbb{E}[\nabla_\theta \log p(X;\boldsymbol{\theta}) \nabla_\theta \log p(X;\boldsymbol{\theta})^\top].
\end{aligned}$$

If $\log p$ is twice differentiable (in $\boldsymbol{\theta}$) the under certain regularity conditions, the Fisher information matrix is equivalent to the expected value of the *negative* Hessian of the log probability,

$$\mathscr{I}(\boldsymbol{\theta}) = -\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}^2 \log p(X;\boldsymbol{\theta})\right]$$

Example: Fisher Information for the Bernoulli Distribution

For a Bernoulli distribution, the log probability and its score were evaluated above,

$$\begin{aligned} \log p(x;\theta) &= x \log \theta + (1-x)\log(1-\theta) \\ \nabla_\theta \log p(x;\theta) &= \frac{x-\theta}{\theta(1-\theta)} \end{aligned}$$

The negative Hessian with respect to θ is,

$$-\nabla_\theta^2 \log p(x;\theta) = \frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}.$$

Example: Fisher Information for the Bernoulli Distribution

Taking the expectation w.r.t. $X \sim p(\cdot;\theta)$ yields,

$$\begin{aligned}\mathscr{I}(\theta) &= -\mathbb{E}\left[\nabla_\theta^2 \log p(x;\theta)\right] \\ &= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} \\ &= \frac{1}{\theta(1-\theta)}.\end{aligned}$$

Interestingly, the inverse Fisher information is the $\mathrm{Var}[X;\theta]$. We'll see that this is a general property of exponential family distributions.

Outline

- Basic Probability Distributions
- Maximum Likelihood Estimation
- **Contingency Tables and Independence**

Contingency Tables

Our table of guesses is an example of a **contingency table**. It represents a sample from a **joint distribution** of two random variables, $X, Y \in \{0, 1, \ldots, 10\}$ indicating the two scores, mod 10.

More generally, let $X \in \{1, \ldots, I\}$ and $Y \in \{1, \ldots, J\}$ be categorical random variables. We represent the joint distribution as an $I \times J$ table,

$$\mathbf{\Pi} = \begin{bmatrix} \pi_{11} & \ldots & \pi_{1J} \\ \vdots & & \vdots \\ \pi_{I1} & \ldots & \pi_{IJ} \end{bmatrix}$$

where

$$\pi_{ij} = \Pr(X = i, Y = j).$$

Contingency Tables

The probabilities must be normalized,

$$1 = \sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} \triangleq \pi_{\bullet\bullet}$$

The **marginal probabilities** are given by,

$$\Pr(X = i) = \sum_{j=1}^{J} \pi_{ij} \triangleq \pi_{i\bullet},$$

$$\Pr(Y = j) = \sum_{i=1}^{I} \pi_{ij} \triangleq \pi_{\bullet j}.$$

Finally, the conditional probabilities are given by Bayes' rule,

$$\Pr(Y = j \mid X = i) = \frac{\Pr(X = i, Y = j)}{\Pr(X = i)} = \frac{\pi_{ij}}{\pi_{i\bullet}} \triangleq \pi_{j|i}$$

Independence

One of the key questions in the analysis of contingency tables is whether X and Y are independent. In particular, they are independent if the joint distribution factors into a product of marginals,

$$X \perp Y \iff \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \; \forall i, j.$$

Equivalently, the variables are independent if the conditionals are *homogeneous*,

$$X \perp Y \iff \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i\bullet}} = \frac{\pi_{i\bullet}\pi_{\bullet j}}{\pi_{i\bullet}} = \pi_{\bullet j} \; \forall i, j.$$

Independence Testing in Multi-Way Tables

Here, we will derive a likelihood ratio test to test for independence in a contingency table. Let

$$\mathscr{H}_0 : \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \quad \forall i, j$$

be our null hypothesis of independence. The null hypothesis imposes a constraint on the set of probabilities $\boldsymbol{\Pi}$. Rather than taking on any value $\boldsymbol{\Pi} \in \Delta_{IJ-1}$, they are constrained to the $\Delta_{I-1} \times \Delta_{J-1}$ subset of probabilities that factor into an outer product of marginal probabilities.

The likelihood ratio test compares the maximum likelihood under the constrained set to the maximum likelihood under the larger space of all probabilities,

$$\lambda = -2 \log \frac{\sup_{\pi_{i\bullet}, \pi_{\bullet j} \in \Delta_{I-1} \times \Delta_{J-1}} p(\boldsymbol{x}; \boldsymbol{\pi}_{i\bullet}\boldsymbol{\pi}_{\bullet j}^\top)}{\sup_{\boldsymbol{\Pi} \in \Delta_{IJ-1}} p(\boldsymbol{x}; \boldsymbol{\Pi})}$$

Independence Testing in Multi-Way Tables

The maximum likelihoods estimates of the constrained model are $\hat{\pi}_{i\bullet} = x_{i\bullet}/x_{\bullet\bullet}$ and $\hat{\pi}_{\bullet j} = x_{\bullet j}/x_{\bullet\bullet}$; under the unconstrained model they are $\hat{\pi}_{ij} = x_{ij}/x_{\bullet\bullet}$. Plugging these estimates in yields,

$$\begin{aligned}\lambda &= -2\log\frac{\prod_{ij}\left(\frac{x_{i\bullet}x_{\bullet j}}{x_{\bullet\bullet}^2}\right)^{x_{ij}}}{\prod_{ij}\left(\frac{x_{ij}}{x_{\bullet\bullet}}\right)^{x_{ij}}} \\ &= -2\sum_{ij} x_{ij}\log\frac{\hat{\mu}_{ij}}{x_{ij}}\end{aligned}$$

where $\hat{\mu}_{ij} = x_{\bullet\bullet}\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j} = x_{i\bullet}x_{\bullet j}/x_{\bullet\bullet}$ is the expected value of X_{ij} under the null hypothesis of independence.

Under the null hypothesis, λ is asymptotically distributed as chi-squared with $(IJ-1)-(I-1)-(J-1) = (I-1)(J-1)$ degrees of freedom,

$$\lambda \sim \chi^2_{(I-1)(J-1)}.$$

Summary

There is a lot more to say about contingency tables that we did not have time to cover.

- Depending on the setting, we can consider various *sampling models* for the contingency table, like Poisson, multinomial, or hypergeometric sampling.
- For two-way tables, there are other statistics for measuring association, like the *relative risk*.
- Likewise, in two-way tables, there are several ways to test for independence, including *Fisher's exact test*, which is based on a *hypergeometric sampling* model.

Next, we'll consider models for capturing relationships between a binary response and several covariates using *logistic regression*.