Probability and Statistics Review

COS 424/524, SML 302: Fundamentals of Machine Learning

Professor Engelhardt

COS 424/524, SML 302

Lecture 2

Introduction: concepts in probability and statistics

Today we will review some essential concepts in probability and statistics

- Probability: definitions
- Independence
- Continuous and discrete distributions: definitions
- Maximum likelihood parameter estimation
- Specific distributions
  - Bernoulli
  - Gaussian
  - Poisson
  - Multinomial

It is essential to be comfortable with these concepts in order to develop, extend, fit, and understand the assumptions in statistical models.

Card problem (from David MacKay)

- There are three cards
  - Red/Red
  - Red/Black
  - Black/Black
- I go through the following process.
  - Close my eyes and pick a card
  - Pick a side at random
  - Show you that side

Suppose I show you red. What is the probability the other side is red too?

Random variable

- A *random variable* is a unique value that is associated with an experimental outcome. Repeated experiments will produce different values of the random variable
- *Probability (Classical definition):* quantifies the long-run frequency that a *random variable* will take a specific value
- *Probability (Bayesian definition):* quantifies our belief in the certainty that a *random variable* will take a specific value

Random variables: examples

Random variables

- The result of a flip of a coin
- The height of someone chosen randomly from a population
- Whether or not you will be diagnosed with type II diabetes
- Number of colleges you applied to
- The temperature in Princeton on a certain day
- The number of times "streetlight" appears in a document

It is useful to think of many feature types as random variables.

Random variables

Random variables can be *discrete* or *continuous*.

Discrete RVs

- Coin flip: $\{H, T\}$
- Number of words in a document: Positive integers $\mathcal{Z}^+ = \{1, 2, 3, \ldots\}$

Continuous RVs

- Height: positive real values $\Re^+ = (0, \infty)$
- Temperature: real values $\Re = (-\infty, \infty)$

We will focus for the first part of this lecture on discrete RVs.

Discrete distributions: Atoms

- An *atom* is a value that a random variable may take.
- Atoms are mutually exclusive: a random variable cannot take on two (atomic) values as the result of one experiment.
- The sum of the probability of a random variable taking on all of the atoms in a *sample space* Ω is one.
- Denote the random variable with a capital letter; denote a realization of the random variable with a lower case letter.

Atoms: examples

- X is a coin flip, x is the value (H or T) of that coin flip.
- X is the roll of a die; x is the value (1, 2, 3, 4, 5, or 6)

Discrete distributions: more formally

Discrete distributions assign a probability to atoms in sample space Ω

Example: let X be the flip of an unfair coin

$$\begin{aligned} P(X = H) &= 0.7 \\ P(X = T) &= 0.3 \end{aligned}$$

- Probabilities over atoms in the sample space sum to one

$$\sum_{x \in \Omega} P(X = x) = 1$$

- Probabilities of *events* are sums over probabilities of atoms.

The probability that a die roll is bigger than 3?

$$P(D > 3) = P(D = 4) + P(D = 5) + P(D = 6)$$

Probability spaces and events

A *probability space* is defined as:

- The sample space Ω
- The set of all events $\mathcal{F}$
- A probability function mapping an event to a probability

Example: roll a fair die

- The RV is $D = d$, the outcome of the rolled fair die
- The atoms in the sample space are $\Omega = \{1, 2, 3, 4, 5, 6\}$
- The events $\mathcal{F}$ are all combinations of atoms $(2^{|\Omega|})$, e.g., $d > 3$
- The probability function maps event to a probability: $P(D > 3) = \frac{1}{2}$

Joint distribution: distribution over collections of RVs

- The *joint distribution* is a distribution over the configuration of all the random variables in a collection.

Example: flip four coins

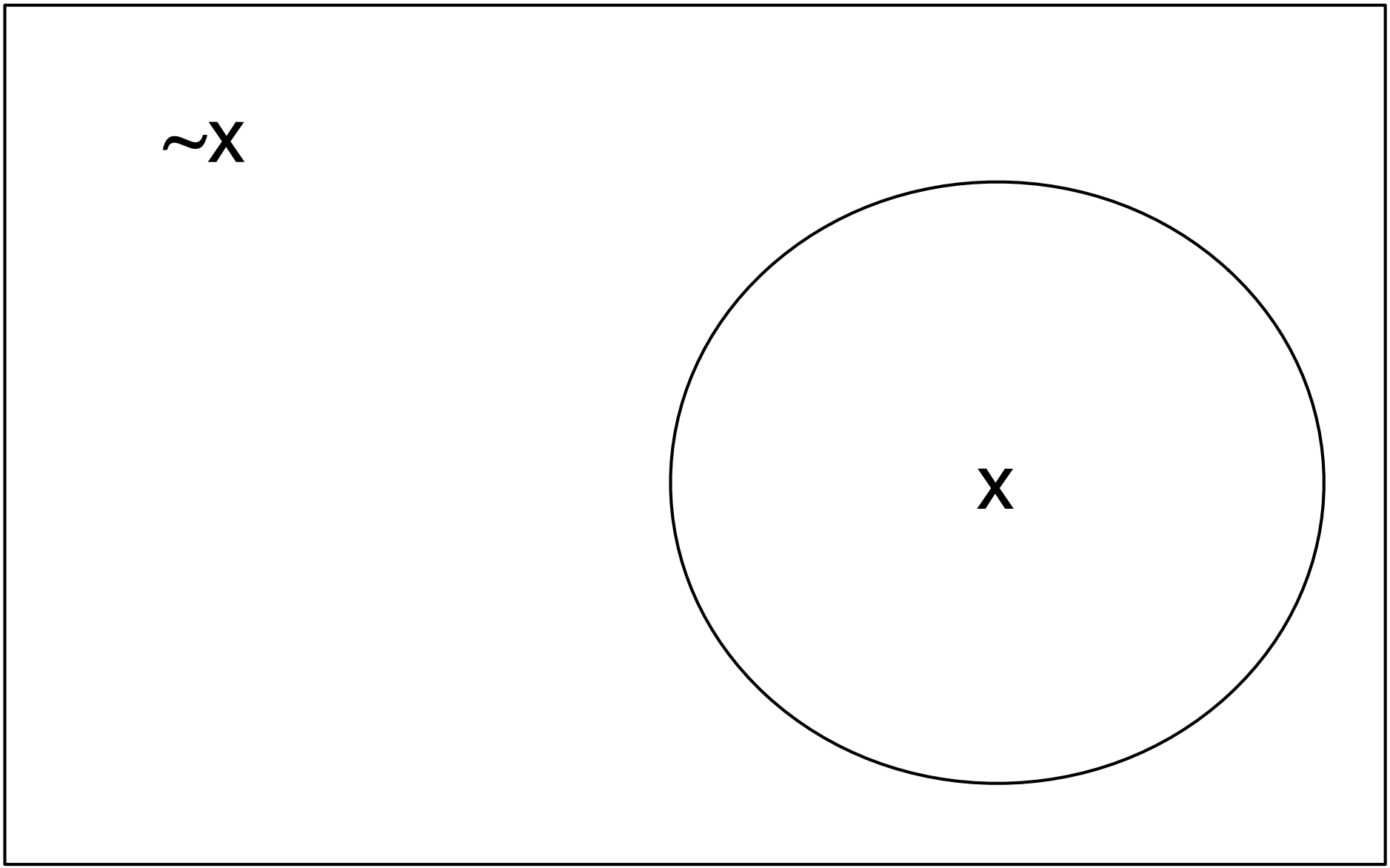
Imagine flipping four fair coins. The joint distribution is over the space of all possible outcomes of the four coins.

$$\begin{aligned} P(HHHH) &= 0.0625 \\ P(HHHT) &= 0.0625 \\ P(HHTH) &= 0.0625 \\ &\dots \end{aligned}$$

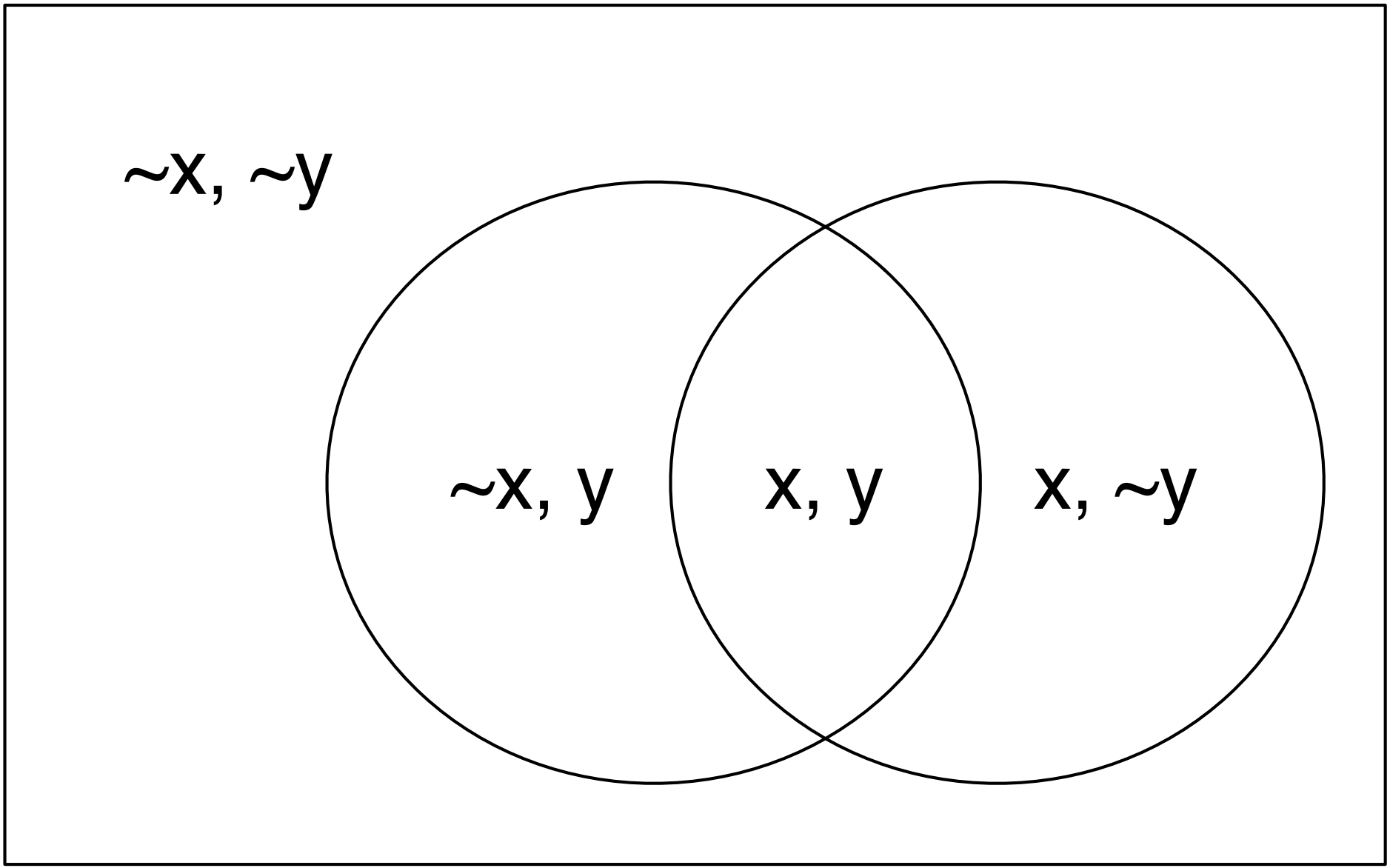
You can think of the outcome of this four-coin flip as a single random variable with a probability space that contains sixteen atoms.

What is the probability of exactly one head?

Visualizing a joint distribution



Visualizing a joint distribution



Conditional distribution

- A *conditional distribution* is the distribution of a random variable given an observation of another random variable.

Example: conditional distribution

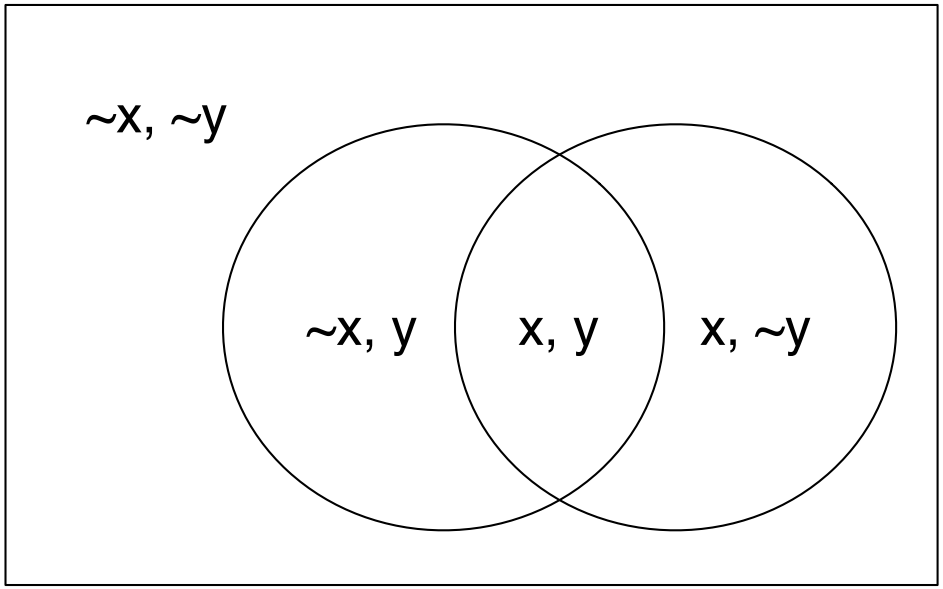
$$\begin{aligned} P(\text{Kids are home}) &= 0.8 \\ P(\text{I listen to Beatles} \mid \text{Kids are home}) &= 0.5 \\ P(\text{I listen to Beatles} \mid \text{Kids are not home}) &= 0.9 \end{aligned}$$

- $P(X = x|Y = y)$ is a different distribution for each value of y

$$\sum_{x \in \Omega_x} P(X = x \mid Y = y) = 1$$

$$\sum_{y \in \Omega_y} P(X = x \mid Y = y) \neq 1 \quad \text{(at least, not necessarily)}$$

Definition of conditional probability



- Conditional probability is defined as:

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)},$$

  which holds when $P(Y = y) > 0$.

- In the Venn diagram, this is the relative probability of $X = x$ in the subspace of the probability space where $Y = y$.

Now we can solve the card problem

Card problem with three cards

There are three cards: Red/Red, Red/Black, Black/Black.
I go through the following process.

- Close my eyes and pick a card
- Pick a side at random
- Show you that side

Suppose I show you red. What is the probability the other side is red too?

- Let X_1 be the random side of the random card I chose
- Let X_2 be the other side of that card
- Compute $P(X_2 = \text{red} \mid X_1 = \text{red})$

$$P(X_2 = \text{red} \mid X_1 = \text{red}) = \frac{P(X_1 = R, X_2 = R)}{P(X_1 = R)}$$

Now we can solve the card problem

Now we can solve the card problem.

- Let X_1 be the random side of the random card I chose
- Let X_2 be the other side of that card
- Compute $P(X_2 = \text{red} \mid X_1 = \text{red})$

$$P(X_2 = \text{red} \mid X_1 = \text{red}) = \frac{P(X_1 = R, X_2 = R)}{P(X_1 = R)}$$

- Numerator is $1/3$: Only one card has two red sides.
- Denominator is $1/2$: Three sides out of six are red.
- So $P(X_2 = \text{red} \mid X_1 = \text{red}) = 2/3$

Is it possible for a conditional probability to be outside of $[0, 1]$?

Gender bias at Berkeley: Conditional probabilities

From the textbook *Statistics* by Freedman, Pisani, & Purves:

> **An observational study on sex bias in admissions was done by the Graduate Division at the University of California, Berkeley.[7] During the study period, there were 8,442 men who applied for admission to graduate school and 4,321 women. About 44% of the men and 35% of the women were admitted. Taking percents adjusts for the difference in numbers of male and female applicants: 44 out of every 100 men were admitted, and 35 out of every 100 women.**

These numbers show that the system is biased ($p = 1.333 \times 10^{-10}$)

Gender bias at Berkeley: Simpson's Paradox

From the textbook *Statistics* by Freedman, Pisani, & Purves:

When the graduate school asked each of the departments to make changes in their admissions policy, all of the departments noted that *none of their admissions numbers were biased against women.*

What important variable are we not incorporating in this comparison?

Gender bias at Berkeley: Condition on department

From the textbook *Statistics* by Freedman, Pisani, & Purves:

Table 2. Admissions data for the graduate programs in the six largest majors at University of California, Berkeley.

	Men		*Women*	
Major	*Number of applicants*	*Percent admitted*	*Number of applicants*	*Percent admitted*
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Note: University policy does not allow these majors to be identified by name.
Source: The Graduate Division, University of California, Berkeley.

Conditioning on the department, the bias is reversed. Why?

Gender bias at Berkeley: Condition on department

From the textbook *Statistics* by Freedman, Pisani, & Purves:

Table 2. Admissions data for the graduate programs in the six largest majors at University of California, Berkeley.

	Men		*Women*	
Major	*Number of applicants*	*Percent admitted*	*Number of applicants*	*Percent admitted*
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Note: University policy does not allow these majors to be identified by name.
Source: The Graduate Division, University of California, Berkeley.

Women are more likely than men to apply for admission in departments with low admission rates.

The chain rule

- The definition of conditional probability lets us derive the *chain rule*
- The *chain rule* factorizes a joint distribution as a product of conditional distributions:

$$\begin{aligned} P(X, Y) &= P(X, Y)\frac{P(Y)}{P(Y)} \\ &= P(X \mid Y)P(Y) \end{aligned}$$

Example: use conditional and marginal to get joint distribution

Let Y be a disease and X be a symptom.

We may know $P(X \mid Y)$ and $P(Y)$ from data.

Use chain rule to find the probability of the disease and the symptom.

The chain rule across n variables

In general, for any set of n variables

$$P(X_1, \ldots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i \mid X_1, \ldots, X_{i-1})$$

Marginalization

- Given a collection of RVs, we might only consider a subset of them.
- Compute $P(X)$ using joint distribution $P(X, Y, Z)$
- Can do this with *marginalization*

$$P(X) = \sum_{y \in \Omega_y} \sum_{z \in \Omega_z} P(X, Y = y, Z = z)$$

Proof.

$$\begin{aligned}
\sum_{y \in \Omega_y} P(X, Y = y) &= \sum_{y \in \Omega_y} P(X) P(Y = y \mid X) \\
&= P(X) \sum_{y \in \Omega_y} P(Y = y \mid X) \\
&= P(X)
\end{aligned}$$

Bayes rule

- From the chain rule and marginalization, we obtain *Bayes rule*.

$$\begin{aligned} P(Y \mid X) &= \frac{P(X,Y)}{P(X)} = \frac{P(X \mid Y)P(Y)}{P(X)} \\ &= \frac{P(X \mid Y)P(Y)}{\sum_y P(X \mid Y = y)P(Y = y)} \end{aligned}$$

Example: Bayes rule

Let Y be a disease and X be a symptom.

From $P(X \mid Y)$, $P(X)$ and $P(Y)$, we can compute $P(Y \mid X)$.

- Bayes rule is useful because we can flip a conditional probability.
- More on the interpretation of Bayes rule in the next lecture

Independence

- Random variables are *independent* if knowing the outcome X does not change the probability of Y:

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

- This means that their joint distribution factorizes:

$$X \perp\!\!\!\perp Y \iff P(X, Y) = P(X)P(Y).$$

- Why? The chain rule

$$\begin{aligned} P(X, Y) &= P(X)P(Y \mid X) \\ &= P(X)P(Y) \end{aligned}$$

Independent RVs: Examples

Examples of independent random variables

- Flipping a coin once / flipping the same coin a second time
- You use an electric toothbrush / blue is your favorite color
- You like pineapple on your pizza / voted for Trump

Examples of not independent random variables

- Registered as a Republican / voted for Trump
- The color of the sky / time of day
- Age / shoe size

Are these independent?

- Rolls from two twenty-sided dice
- Roll three dice to compute two random variables $(D_1 + D_2, D_2 + D_3)$
- # enrolled students and the temperature outside today
- # students taking walks outside and the temperature outside today

Two coins

- Suppose we have two coins, one biased and one fair,

$$P(C_1 = H) = 0.5 \quad P(C_2 = H) = 0.7.$$

- We choose one of the coins at random $Z \in \{1, 2\}$, flip C_Z twice, and record the outcome (X, Y).
- Are X and Y independent?

Two coins

- Suppose we have two coins, one biased and one fair,

$$P(C_1 = H) = 0.5 \quad P(C_2 = H) = 0.7.$$

- We choose one of the coins at random $Z \in \{1, 2\}$, flip C_Z twice, and record the outcome (X, Y).

- If the answer is not straightforward, consider $P(C_2 = H) = 0.99$

Two coins

- Suppose we have two coins, one biased and one fair,

$$P(C_1 = H) = 0.5 \quad P(C_2 = H) = 0.7.$$

- We choose one of the coins at random $Z \in \{1, 2\}$, flip C_Z twice, and record the outcome (X, Y).

What if we knew which coin was flipped Z?

What if the coins have equal probability of heads?

Conditional independence

- X and Y are *conditionally independent* given Z:

$$P(Y \mid X, Z = z) = P(Y \mid Z = z)$$

for all possible values of z.

- Again, this implies a factorization

$$X \perp\!\!\!\perp Y \mid Z \iff P(X, Y \mid Z = z) = P(X \mid Z = z)P(Y \mid Z = z),$$

for all possible values of z.

How would you prove this?

Continuous random variables

- Random variables can be continuous.
- We need a *density* $p(x)$, which *integrates* to one.
  If $x \in \Re$ then

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

- Probabilities are integrals over smaller intervals. E.g.,

$$P(X \in (-2.4, 6.5)) = \int_{-2.4}^{6.5} p(x)dx$$

Example: Gaussian distribution

Continuous distribution: Gaussian

- The Gaussian (or normal) distribution is a continuous distribution, meaning that its *support* is on continuous numbers.

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- The density of a point x is proportional to the negative exponentiated half distance to μ scaled by σ^2.
- Parameters: μ is called the *mean*; σ^2 is called the *variance*.

Notation for random variables

Discrete RVs

- p denotes the *probability mass function*, which is the same as the probability of atoms P.
- We can use P and p interchangeably for discrete distributions.

Continuous RVs

- p is the *probability density function*
- the probability of any single value is always zero: $P(X = x) = 0$.
- We cannot use P and p interchangeably for continuous distributions.

This is an unpleasant detail, but mathematically important.

Expectation

- Consider a function $f(\cdot)$ of a discrete random variable X. (Note: $f(X)$ is also a random variable.)
- The expectation is a weighted average of $f(\cdot)$, where the weight of each atom is $p(x)$,

$$\mathrm{E}[f(X)] = \sum_{x \in \Omega} p(x) f(x)$$

- In the continuous case, the expectation is an integral

$$\mathrm{E}[f(X)] = \int_{\Omega} p(x) f(x) dx$$

Conditional expectation

- The conditional expectation is defined similarly

$$\mathrm{E}[f(X) \mid Y = y] = \sum_{x \in \Omega} p(x \mid y) f(x)$$

Examples: conditional expectations

- Given someone's height Y, what is their expected shoe size X?
- Given someone's zip code, what is the expected number of dollars they will spend on your website?

- $\mathrm{E}[f(X) \mid Y]$ is a (function of) random variable Y.

Probability models

Probability distributions are simple *models* of data that we observe.

- Assume that data are *generated* from a specific distribution.
- *Infer* the parameters of that distribution from the observed data.
- *Interpret* those properties in terms of properties of the underlying observed data.

Probability models, examples

Examples of inferences we can make about specific data sets using a simple distribution:

Inference	Observation
the bias of a coin	1000 coin flips
the average height of a student	190 students
the chance that a politician will win a primary	1200 people in Iowa
the proportion of gold in a mountain	20 cubic meters of the mountain
the number of bacteria species in our body	4 samples of the gut
the evolutionary rate at which DNA mutates	34 completely sequenced mammals

What distribution should be used in order to make each inference?

Independent and identically distributed random variables

- Independent and identically distributed (IID) random variables are:
  1. Independent
  2. Identically distributed
- If we repeatedly flip the same coin n times and record the outcome, then $X_1, \ldots, X_n$ are IID.
- The IID assumption is useful in data analysis.
- But, while the IID assumption is useful, it rarely holds in practice.

Statistical terminology: data

On the data side:

- *Sample* (n) is n IID draws of a specific random variable X
- *Features* (p) is the dimension of random variable X

Examples

- Emails: n separate emails, p word counts from a dictionary of length p
- Netflix: n users, p movies
- Microarray data: n samples, p genes

Statistical terminology: models

On the statistical model side:

- *Parameters* are values that define (or *index*) a distribution
  - scale with the number of features $O(p)$
- *Latent variables* are features that cannot be directly observed
  - scale with the number of samples $O(n)$
- *Observed variables* are features that are observed
  - may be thought of as a $n \times p$ matrix

Example: an email filter where features are dictionary words

- parameters are the frequency of each word for *spam*, *not spam*
- latent variables are assignments of unlabeled email to *spam* or *not spam*
- observed variables are the dictionary word counts for each sample

What is a parameter?

Parameters are values that *index* a distribution.

Bernoulli parameters

A coin flip is a *Bernoulli* distribution.

The Bernoulli parameter (*bias*) is the probability of a H (refer to H as 1).

$$p(x \mid \pi) = \pi^{\mathbb{1}[x=1]}(1-\pi)^{\mathbb{1}[x=0]},$$

where $\mathbb{1}[\cdot]$ is an *indicator function*, which is 1 when its argument is true and 0 otherwise.

Changing π leads to different Bernoulli distributions.

The likelihood function

The data *likelihood function* is the probability of the observed data X given the model parameters θ:

$$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$$

Likelihood of a sequence of coin flips

- Suppose we flip a coin n times and record the outcomes.
- Further, suppose *we think* that the probability of heads is π. (We do not yet care about estimating the true π.)
- Given π, the likelihood, or probability of an observed sequence, is

$$p(x_1, \ldots, x_n \mid \pi) = \prod_{i=1}^{n} \pi^{\mathbb{1}[x_i=1]}(1-\pi)^{\mathbb{1}[x_i=0]}$$

Why can I multiply likelihoods across the samples?

The log likelihood

Take the log of the likelihood function; this is the *log likelihood function*.

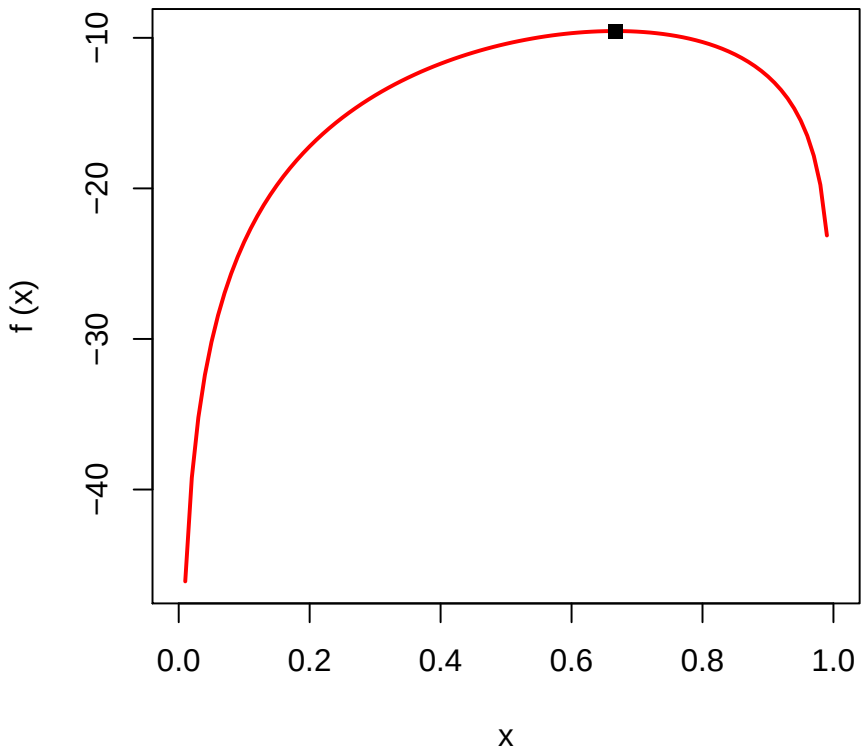
$$\begin{aligned}
\ell\ell(\pi; \mathbf{x}) &= \log p(\mathbf{x} \mid \pi) \\
&= \log \prod_{i=1}^{n} \pi^{\mathbb{1}[x_i=1]} (1-\pi)^{\mathbb{1}[x_i=0]} \\
&= \sum_{i=1}^{n} \mathbb{1}[x_i = 1] \log \pi + \mathbb{1}[x_i = 0] \log(1-\pi)
\end{aligned}$$

The log likelihood is the objective in an optimization problem:

What is the value of the parameter that maximizes the log likelihood?

Do the log likelihood and the likelihood have the same optima?

Bernoulli log likelihood



- We observe ten Hs, five Ts.
- The value of π that maximizes the log likelihood is $2/3$.

How many optima does this function have?

The maximum likelihood estimate

The *maximum likelihood estimate* (MLE) of a parameter is the value of that parameter that maximizes the log likelihood.

Example: MLE for Bernoulli parameter

MLE estimate of π is the proportion of heads.

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[x_i = 1]$$

In a sense, the MLE is the value that best explains our observations.

The MLE is consistent

- The MLE of a parameter is *consistent*: as the number of samples goes to infinity, the parameter estimate converges to the true parameter.

Consistency example

- Say we have n samples from a Bernoulli with true bias π^*.
- Estimate the parameter π from $x_1, \ldots x_n$ with the MLE $\hat{\pi}_{MLE}$.
- Then,

$$\lim_{n \to \infty} \hat{\pi}_{MLE} = \pi^*$$

- This is usually a good thing. It lets us sleep at night.

When will the MLE fail us?

How to compute the MLE for a given distribution?

We want to solve the following optimization problem:

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \ell\ell(\theta).$$

- Often, a single maximum or a closed form solution does not exist

When the log likelihood is convex, we can:

- Take the derivative of the log likelihood function with respect to θ
- Set this derivative to zero
- Solve for θ.

Example: MLE for the Gaussian mean

Example: Gaussian mean

The log likelihood is

$$\ell\ell(\mu, \sigma) = -\frac{1}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

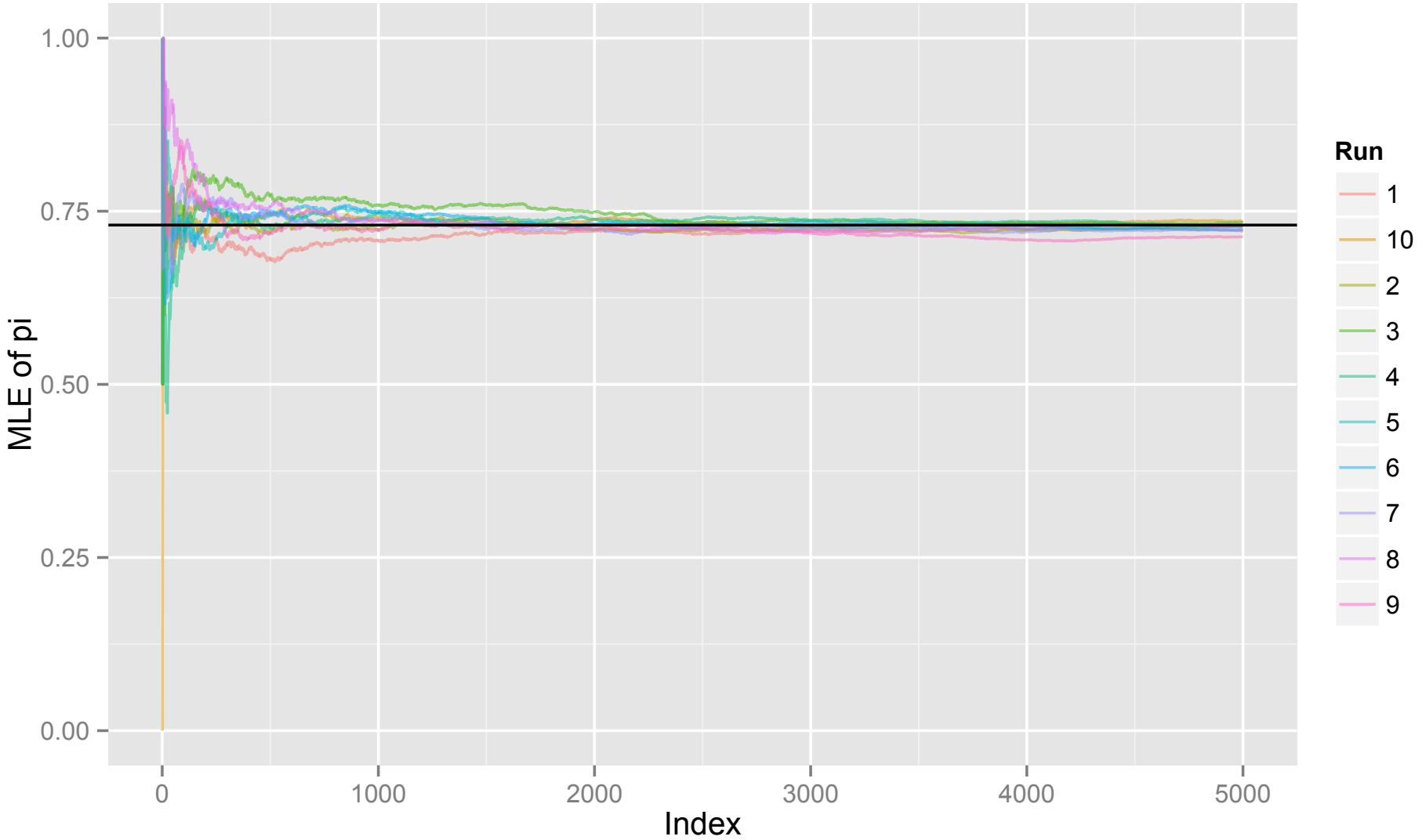
Take the derivative with respect to μ, set to 0:

$$\begin{aligned} \frac{\partial \ell\ell(\mu, \sigma)}{\partial \mu} &\rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \\ \sum_{i=1}^{n}\mu &= \sum_{i=1}^{n} x_i \\ \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ the sample mean} \end{aligned}$$

Example: 5000 coin flips

1 1 0 1 1 1 1 0 0 1 0 0 1 1 1 0 0 0 0 0 1 0 0 1 0 0 1 1 1 0 1 0 1 0 0 0 0 1
0 1 0 1 1 1 1 0 0 0 1 1 0 1 1 1 1 1 1 0 1 1 0 1 0 1 1 1 0 0 0 1 1 1 1 1 0 1
1 1 1 1 1 0 1 1 0 1 1 1 1 0 0 1 1 1 0 0 1 0 1 0 0 0 0 1 0 0 1 1 1 0 1 1 1 0
1 1 1 0 1 1 0 0 0 1 1 1 1 1 1 0 1 1 1 0 0 1 1 1 1 1 0 0 0 0 1 0 1 1 1 1 1 1
1 0 0 1 1 1 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 1 1 0 1 1 0 0 1 1 1 1 1 1 0 1
0 0 1 0 0 1 0 0 1 1 0 0 1 0 0 0 1 1 1 0 1 1 0 0 0 1 1 0 1 1 1 0 1 1 0 0 1 0
1 0 1 1 1 1 1 0 0 0 0 1 1 0 0 1 0 1 0 0 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 1 0
1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 0 0 1 0 0 1 1 1
0 0 1 1 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 1 1 0 1 1 1 0 1 1 1 1 0 1 1 0 0 0 0 1
1 1 0 1 0 1 0 1 0 1 1 0 1 0 0 1 1 1 0 0 1 1 1 0 1 0 1 0 1 1 0 1 1 1 1 1 1 0 0
0 1 1 1 1 1 1 0 1 1 1 0 0 0 1 0 1 0 0 1 1 1 1 1 0 1 0 1 0 0 1 1 0 1 1 1 1 0
0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 0 1 0 1 1 1 1 1 1 0 0 1 1 0 1 1 1 0 0 1 0 0
1 0 1 1 1 1 0 0 1 1 1 1 1 0 1 1 0 0 1 0 1 1 0 1 1 1 1 0 0 0 0 1 0 0 1 1 1 1
0 0 0 0 1 0 0 1 1 1 1 1 0 1 1 0 1 1 1 1 1 0 1 1 0 1 0 1 1 1 0 0 1 0 1 1 1 1
1 1 1 1 0 1 0 1 1 0 0 1 0 1 1 1 0 0 0 1 1 1 0 1 1 1 0 0 0 1 1 1...

Consistency of the MLE example



Important distributions that we will discuss a lot

Let's talk briefly about a few distributions.

- Bernoulli
- Multinomial
- Poisson
- Gaussian

The question I hope to answer in this discussion is:

When I analyze a data set, what is the most appropriate distribution to select to model specific features?

Important distributions to know: Bernoulli

A Bernoulli distribution models a binary random variable

- *Support*: $x \in \{0, 1\}$
- *Parameter*: $\pi \in [0, 1]$ (probability of heads, or *bias*)
- *Probability mass function*:

$$p(x \mid \pi) = \pi^{\mathbb{1}[x=1]}(1-\pi)^{\mathbb{1}[x=0]}$$

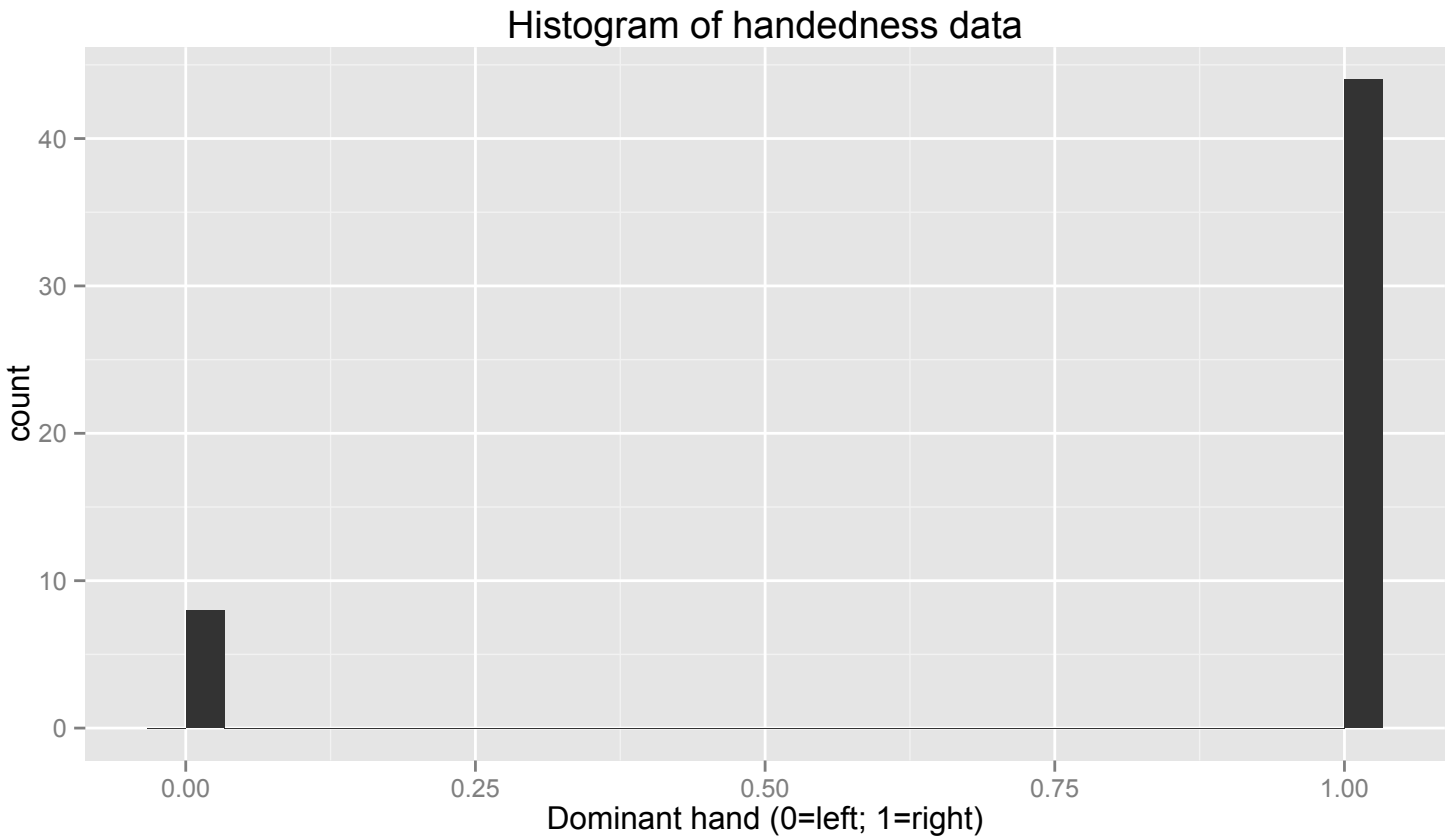
- *MLE estimates of* π:

$$\hat{\pi} = \frac{n_1}{n_0 + n_1},$$

  where n_1 is the number of 1s and n_0 is the number of 0s in n samples.
- *Conjugate prior for* π: beta distribution

The *binomial distribution* models m draws from a Bernoulli distribution.

Modeling dominant hand using a Bernoulli distribution



MLE estimates tell us the right-handed bias is 0.85 in this class.

Important distributions to know: Multinomial

- *Support*: Vectors of counts of m independent draws from one of K categories $[x_1, \ldots, x_K]$
- *Parameter*: θ, $\theta_k \geq 0$, $1 = \sum_{k=1}^{K} \theta_k$
- *Probability mass function*:

$$p(x \mid \theta) = \frac{m!}{x_1! \ldots x_K!} \theta_1^{x_1} \ldots \theta_K^{x_K}$$

- *MLE estimates of* θ:

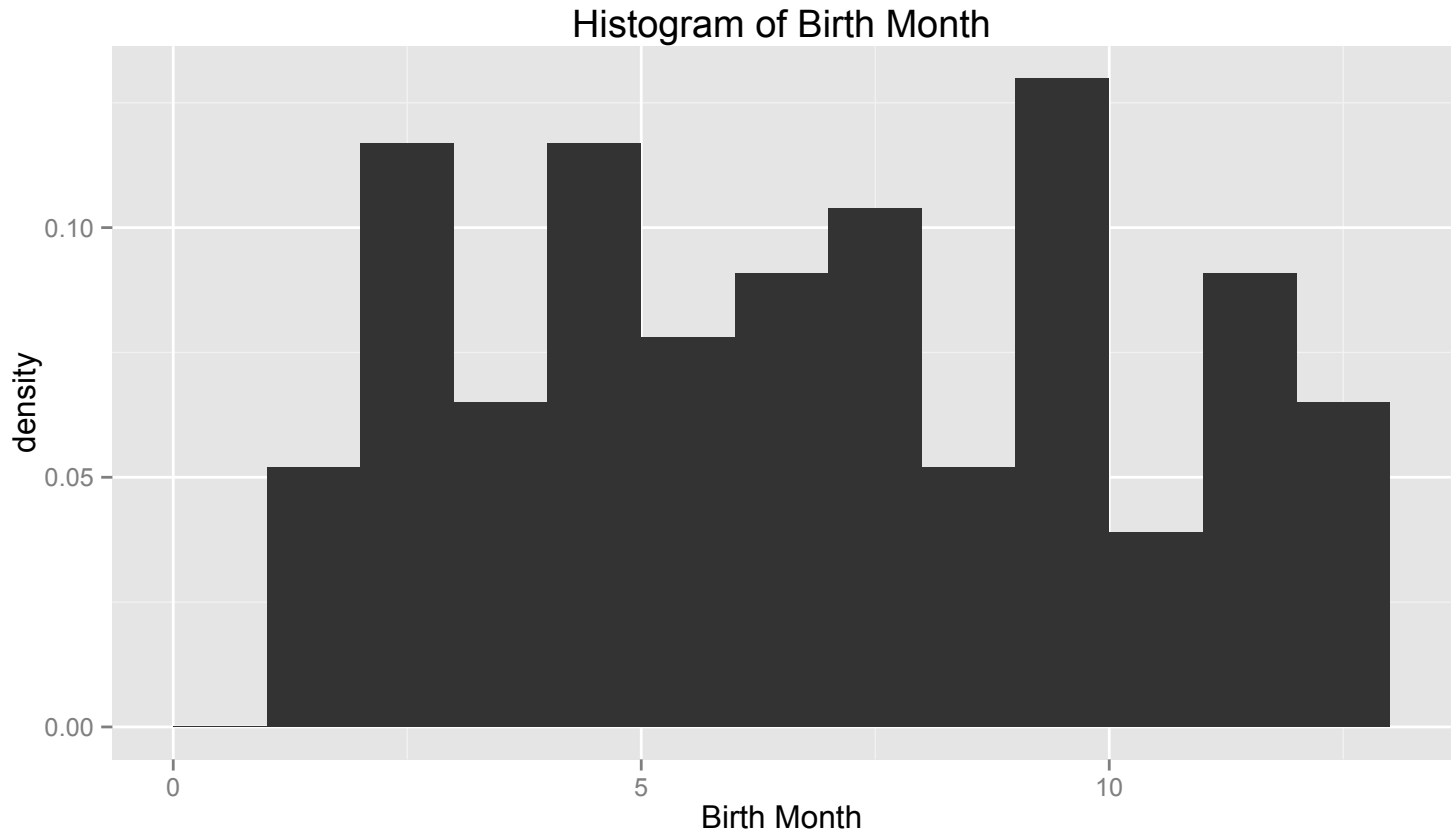
$$\hat{\theta} = \left[\frac{x_1}{m}, \ldots, \frac{x_K}{m} \right],$$

  where x_k is the count of observations in category k
- *Conjugate prior for* θ: Dirichlet distribution

Data that have been modeled using a Multinomial

- Rolls of a die
- Votes for candidates in an election
- Word frequencies in documents
- Movie/book/song ratings

Modeling birth month using a multinomial distribution



The MLE estimates say you are three times more likely to be born in September than in October.

Important distributions to know: Poisson

The Poisson distribution naturally estimates count data

- *Support*: non-negative integers $\{0\} \bigcup \mathcal{Z}^+$
- *Parameter*: $\lambda \in \Re^+$ (mean and variance)
- *Probability mass function*:

$$p(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- *MLE estimates of* λ (the empirical mean):

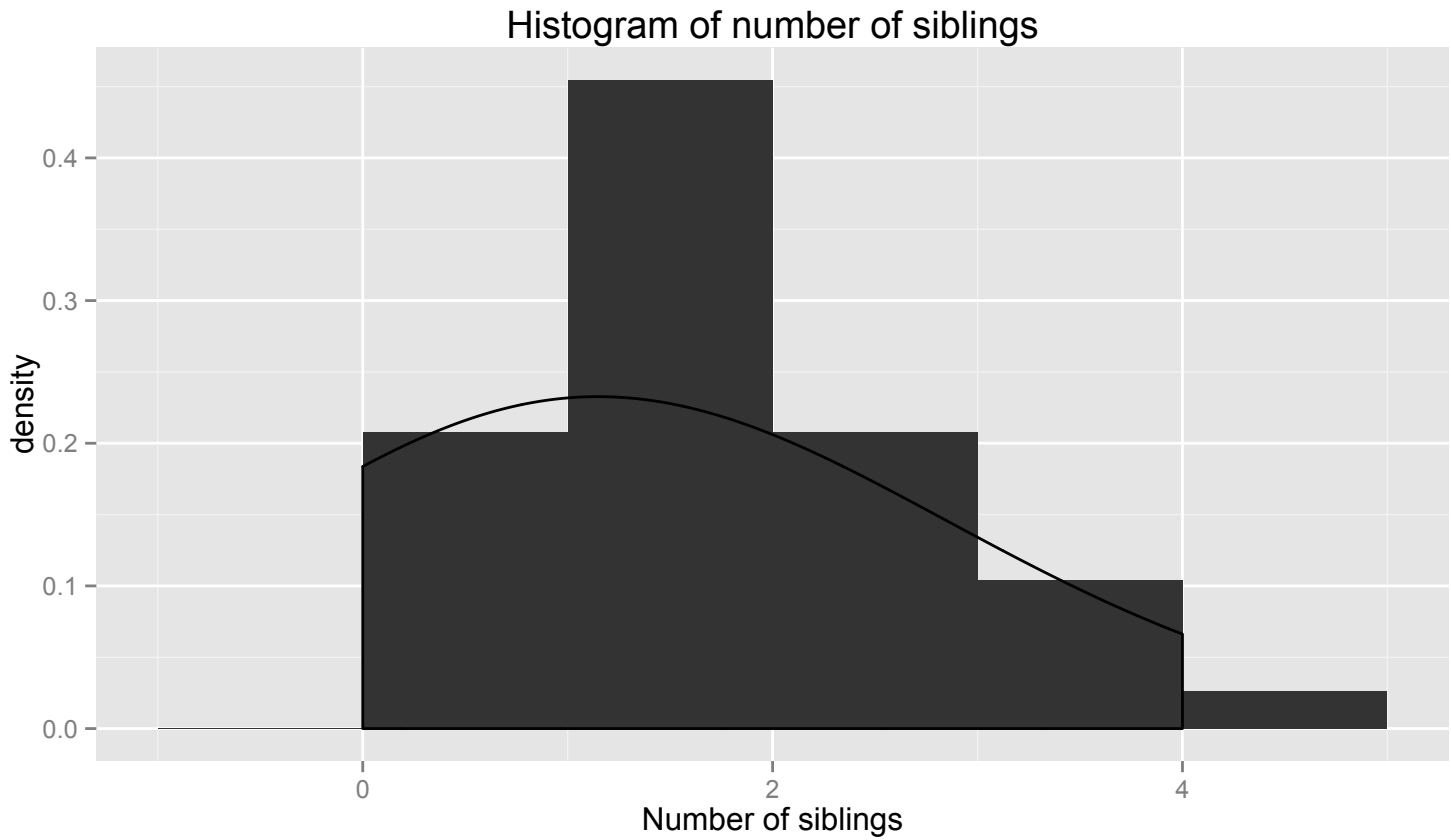
$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- *Conjugate prior for* λ: gamma distribution

Data that have been modeled using a Poisson

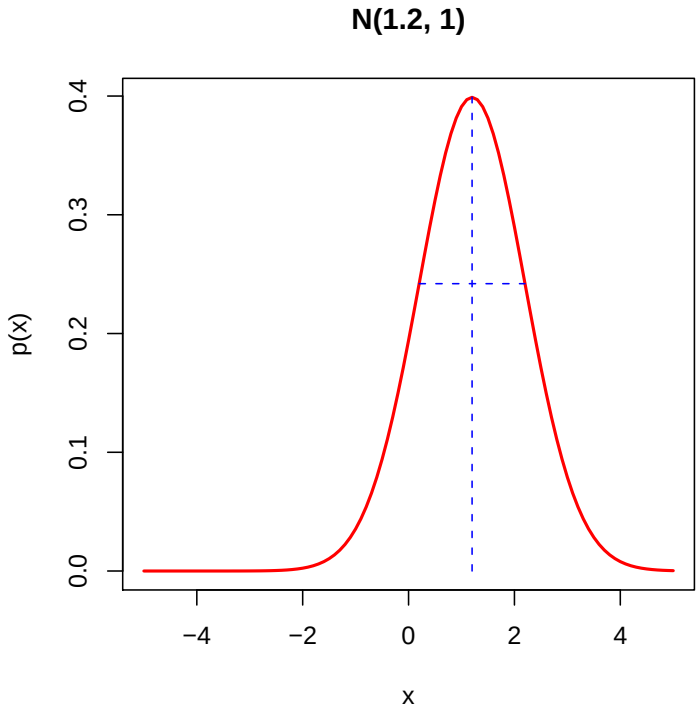
- The number of soldiers killed by horse-kicks each year in each corps in the Prussian cavalry [Bortkiewicz]
- The number of yeast cells used when brewing Guinness beer [Gosset]
- The number of phone calls arriving at a call center within a minute [Erlang]
- The number of goals in sports involving two competing teams
- The number of jumps in a stock price in a given time interval
- Under an assumption of homogeneity, the number of times a web server is accessed per minute
- The number of mutations in a given stretch of DNA after a certain amount of radiation
- The proportion of cells that will be infected at a given multiplicity of infection
- The arrival of photons on a pixel circuit at a given illumination and over a given time period
- The targeting of V-1 flying bombs on London during World War II
- The counts of prime numbers in short intervals obey a Poisson distribution provided a certain version of an unproved conjecture of Hardy and Littlewood is true [Gallagher]

Modeling number of siblings using a Poisson distribution



The empirical mean number of siblings: $\hat{\lambda} = 1.3$; empirical variance is 1.

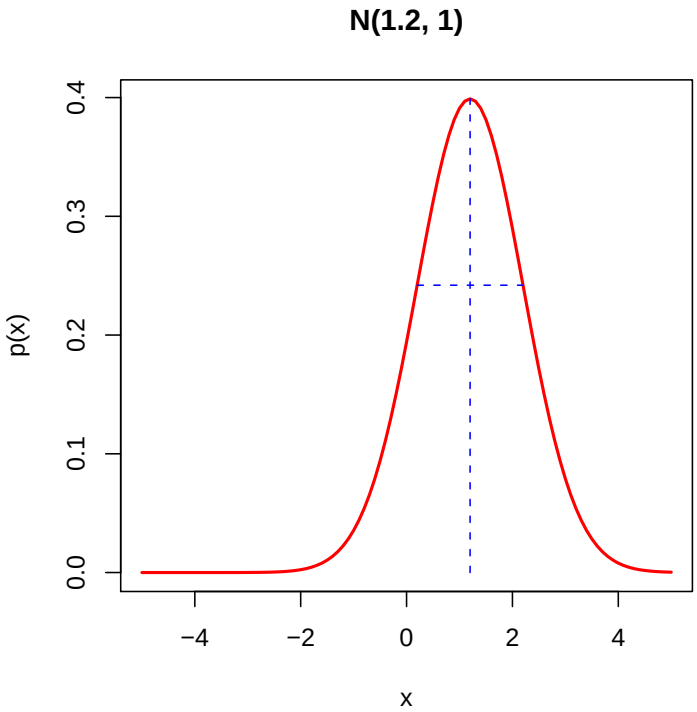
Gaussian density



- The mean μ controls the location of the center of the distribution:

$$\begin{aligned} x &\sim \mathcal{N}(\mu, \sigma^2) \\ x - \mu &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

Gaussian density



- The variance σ^2 controls the spread of the distribution.
- The standard deviation σ is the square root of variance.

$$
\begin{aligned}
p(\mu - \sigma \leq x \leq \mu + \sigma) &\approx 0.6827 \\
p(\mu - 2\sigma \leq x \leq \mu + 2\sigma) &\approx 0.9545 \\
p(\mu - 3\sigma \leq x \leq \mu + 3\sigma) &\approx 0.9973
\end{aligned}
$$

Important distributions to know: Gaussian

- *Support*: $x \in \Re$
- *Parameters*: μ, the mean, and σ^2, the variance
- *Probability density function*:

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- *MLE estimates of* μ (the empirical mean):

$$\hat{\mu} = \overline{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$$

- *MLE estimates of* σ^2 (the empirical variance):

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- *Conjugate prior for* μ: Gaussian distribution
- *Conjugate prior for* σ^2: Inverse gamma distribution

Data modeled using a Gaussian distribution

The *central limit theorem (CLT)*: under certain conditions, the arithmetic mean of a sufficiently large number of observations of independent random variables will be approximately normally distributed regardless of the underlying distribution

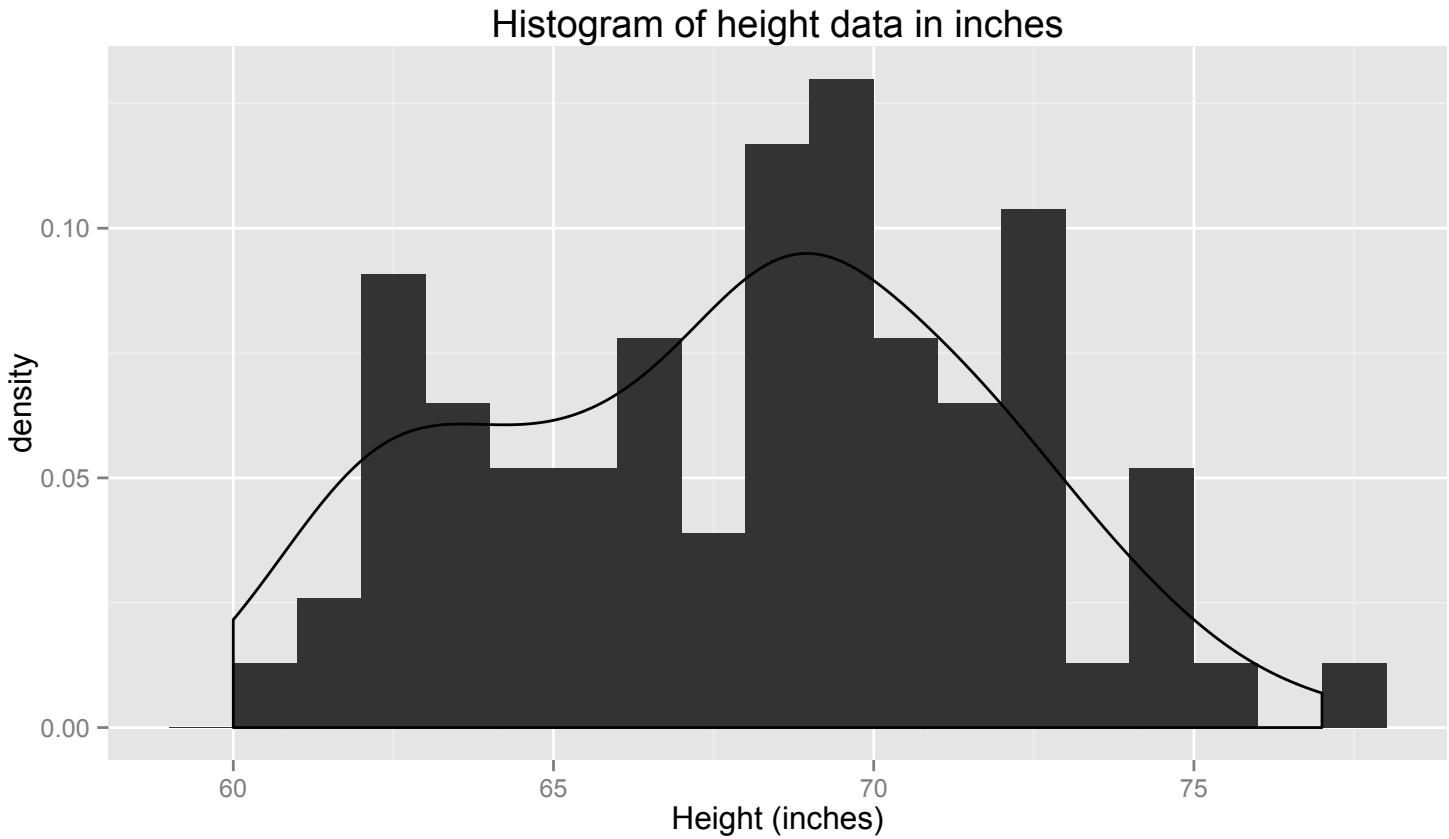
Data modeled using a Gaussian distribution

- In biology, the log of random variables tend to have a normal distribution (after separating male/female subpopulations) including:
  - Measures of size (length, height, skin area, weight)
  - The length of inert appendages (hair, claws, nails, teeth)
  - Physiological measurements, such as blood pressure of adult humans
- In finance, in particular the Black-Scholes model, changes in the logarithm of exchange rates, price indices, and stock market indices are assumed normal (these variables behave like compound interest, not like simple interest, and so are multiplicative)
- In hydrology, the distribution of long duration river discharge or rainfall, e.g., monthly and yearly totals

Data modeled using a Gaussian distribution

- Measurement errors in experiments: using the normal distribution produces the most conservative predictions possible given only knowledge about the mean and variance of the errors
- In standardized testing, results are modeled using a Gaussian. E.g., the SAT's traditional range of 200–800 is based on a normal distribution with a mean of 500 and a standard deviation of 100.
- Percentile ranks ("percentiles" or "quantiles"), normal curve equivalents, and z-scores.
- Bell curve grading assigns relative grades based on a normal distribution of scores

Modeling height data using a Gaussian distribution



$\hat{\mu}_{MLE} = 67.8\text{in (5 ft 7.8 in)}, \hat{\sigma}^2_{MLE} = 15.4$

What would be a better way to model these data?

Model pitfalls

- What's wrong with modeling height with a Gaussian distribution?
  - Assigns positive probability to numbers < 0 and > 100
  - Ignores important biological covariates (i.e., male/female)
  - The data do not look like they came from a Gaussian
- "All models are wrong. Some models are useful." (G. Box)

More interesting statistical models

We will extend these distributions to more sophisticated models using the graphical model framework. We will see the following models in this class:

- Naive Bayes classification
- Linear regression and logistic regression
- Generalized linear models
- Hidden variables, mixture models, and the EM algorithm
- Factor analysis and principal component analysis
- Sequential models

Additional resources

- *Machine Learning: A Probabilistic Approach* (Chapter 2)
- Michael Lavine, *Introduction to Statistical Thought* (an introductory statistical textbook with plenty of R examples, and it's online too)
- Chris Bishop, *Pattern Recognition and Machine Learning* (Ch 1 & 2)
- (video) Sam Roweis: *Machine Learning, Probability and Graphical Models, Part 1*
- (video) Michael Jordan: *Bayesian or Frequentist: Which Are You?*
- `wikipedia` (much of the material in today's lecture is available on wikipedia)