Introduction

Jingchen (Monika) Hu

Vassar College

MATH 347 Bayesian Statistics

Outline

Outline

General info

Instructor:	Jingchen (Monika) Hu - jihu@vassar.edu RH 403
Lecture:	Tuesdays and Thursdays 1:30-2:45pm OLB 105
Lab:	Some lectures will be used as labs.
Office hours:	Monday 4:00 - 5:00pm, Wednesday 10:30-11:30am & 1:30-2:30pm, or by appointment. Separate online office hours for remote students.
TAs:	Dahlia Forte, Lucas Krishan, and Henrik Olsson

Required materials

Prerequisite: MATH 220 Multivariable Calculus , MATH 221 Linear Algebra and MATH 241 Probability.

Textbook: First Course in Bayesian Statistical Methods, by Hoff, P. (2010), Springer

Software: We will use the software R/RStudio for labs and project. Download R from *http://www.r-project.org/* and RStudio, from *https://www.rstudio.com/*.

Webpage: Vassar's Moodle. It is your responsibility to check the site for homework, readings, and announcements. Also, check the tentative schedule (Google Docs) frequently for updates.

Course topics

This class is divided into three parts.

- Inference: Bayes theorem, conjugate prior, posterior distribution, HPD interval, predictive distribution, etc. [Hoff Chapters 1, 2, 3, and 5]

Course topics

This class is divided into three parts.

- Inference: Bayes theorem, conjugate prior, posterior distribution, HPD interval, predictive distribution, etc. [Hoff Chapters 1, 2, 3, and 5]
- Computation:
  - Monte Carlo approximation [Hoff Chapter 4]
  - Markov chain Monte Carlo (MCMC), Gibbs sampler, Metropolis-Hastings algorithm, MCMC diagnostics [Hoff Chapters 6, and 10]
  - JAGS (R packages) for implementing MCMC [Additional material]

Course topics

This class is divided into three parts.

- Inference: Bayes theorem, conjugate prior, posterior distribution, HPD interval, predictive distribution, etc. [Hoff Chapters 1, 2, 3, and 5]
- Computation:
  - Monte Carlo approximation [Hoff Chapter 4]
  - Markov chain Monte Carlo (MCMC), Gibbs sampler, Metropolis-Hastings algorithm, MCMC diagnostics [Hoff Chapters 6, and 10]
  - JAGS (R packages) for implementing MCMC [Additional material]
- Applications: Bayesian hierarchical modeling, Bayesian linear regression, latent class modeling, Bayesian time series modeling, Bayesian cognitive modeling etc. [Hoff Chapters 8, and 9, and additional material]

Course topics

This class is divided into three parts.

- Inference: Bayes theorem, conjugate prior, posterior distribution, HPD interval, predictive distribution, etc. [Hoff Chapters 1, 2, 3, and 5]
- Computation:
  - Monte Carlo approximation [Hoff Chapter 4]
  - Markov chain Monte Carlo (MCMC), Gibbs sampler, Metropolis-Hastings algorithm, MCMC diagnostics [Hoff Chapters 6, and 10]
  - JAGS (R packages) for implementing MCMC [Additional material]
- Applications: Bayesian hierarchical modeling, Bayesian linear regression, latent class modeling, Bayesian time series modeling, Bayesian cognitive modeling etc. [Hoff Chapters 8, and 9, and additional material]

No strict boundaries between topics - they depend on each other.

Online/hybrid teaching and learning

- *Liberal Arts Collaborative for Digital Innovation* (LACOL) 10 schools
- Course Share Project
  - Goal: enrich upper level math/stats offering across 10 schools

Online/hybrid teaching and learning

- *Liberal Arts Collaborative for Digital Innovation* (LACOL) 10 schools
- Course Share Project
  - Goal: enrich upper level math/stats offering across 10 schools
  - 2019 - 2020 Academic Year
    - ★ Fall 2019: Steve Miller (Williams College) Operations Research
    - ★ Fall 2019: Monika Hu (Vassar College) Bayesian Statistics

Online/hybrid teaching and learning

- *Liberal Arts Collaborative for Digital Innovation* (LACOL) 10 schools
- Course Share Project
  - Goal: enrich upper level math/stats offering across 10 schools
  - 2019 - 2020 Academic Year
    - ★ Fall 2019: Steve Miller (Williams College) Operations Research
    - ★ Fall 2019: Monika Hu (Vassar College) Bayesian Statistics
- Our course MATH 347 Bayesian Statistics:
  - Fall 2017: 16 Vassar students + 1 Swarthmore student
  - Spring 2019: 16 Vassar students + 1 Amherst + 1 Carleton + 4 Swarthmore students
  - Currently: 16 Vassar students + 1 Amherst + 1 Swarthmore students
    - ★ Lectures will be synchronized online in real time and recorded for remote students; online office hours for remote students
    - ★ Group work across campuses

Online/hybrid teaching and learning cont'd

- Pre-course survey - please fill out; there will be a post-course survey

Online/hybrid teaching and learning cont'd

- Pre-course survey - please fill out; there will be a post-course survey
- My experience:
  - Some learning material are suitable for video form (e.g. guest lecture videos, recaps, R demo)
  - Moodle discussion forum is great to foster a learning community
  - Wish to have more interaction among Vassar and remote students (case studies, projects)
  - Wish to have more interaction among Vassar students: every week please sit with a different neighbor!

Online/hybrid teaching and learning cont'd

- Pre-course survey - please fill out; there will be a post-course survey
- My experience:
  - Some learning material are suitable for video form (e.g. guest lecture videos, recaps, R demo)
  - Moodle discussion forum is great to foster a learning community
  - Wish to have more interaction among Vassar and remote students (case studies, projects)
  - Wish to have more interaction among Vassar students: every week please sit with a different neighbor!
- Previous students' advice/comments:
  - "Try and take the course with at least one other local student, so you have someone to reach out to locally if necessary."
  - "If you are a local student, it will not make a big difference."
  - "Rewatch the youtube lecture videos, go to office hours, and get an early start on the project."
  - "The textbook was very different than in class lectures."

Online/hybrid teaching and learning cont'd

- Responses to post-survey "Do you re-watch the lecture recordings? How often? For what reasons?"
  - "Yes. Rewarded every lecture around twice for Studying" (Vassar)
  - "Yes for revising" (Remote)
  - "Yes, once to learn the material" (Remote)
  - "Yes, when studying for tests, and to learn something that wasn't clear to me the first time I heard it" (Vassar)
  - "Yes, to get explanations for concepts that I was using in my project" (Remote)
  - "I re-watched lectures when doing assignments, and preparing for the exams." (Remote)
  - "I re-watched them usually when trying to do labs and homeworks. And then reviewing key concepts before exams " (Vassar)
  - "Yes, skimmed through the videos to cover things that I missed / didn't have time to write down during class." (Vassar)
  - "Yes, I did rewatch them, mostly to study before exams but also when I had homework questions as well." (Vassar)
  - "Yes. Once every week or 2 weeks. For revision or absence." (Vassar)

Course components

- The three parts of material: inference, computation, applications
- One extra part hidden: review of material from prerequisite courses (e.g. calculus, linear algebra, and probability)

Course components

- The three parts of material: inference, computation, applications
- One extra part hidden: review of material from prerequisite courses (e.g. calculus, linear algebra, and probability)
- Each part is a combination of a selection of the following: readings, lectures, labs, homework, discussions (in-class and online), case study, and project

Course components

- The three parts of material: inference, computation, applications
- One extra part hidden: review of material from prerequisite courses (e.g. calculus, linear algebra, and probability)
- Each part is a combination of a selection of the following: readings, lectures, labs, homework, discussions (in-class and online), case study, and project
- The course project (individual or in pair; cross-campus collaboration is highly encouraged!) is a final product involving inference, computation, and applications

Course project

- Apply the Bayesian methods you've learned in this course and/or explore new Bayesian methods to solve research question(s) of your choice
- Present at a poster session (last class), with a 2-min intro video

Course project

- Apply the Bayesian methods you've learned in this course and/or explore new Bayesian methods to solve research question(s) of your choice
- Present at a poster session (last class), with a 2-min intro video
- Examples of projects from Spring 2019
  - Earnings and Sexual Orientation link
  - Bayes by Backprop: Weight Uncertainty in Neural Networks link
  - Bayesian Word Embeddings link
  - Full list of projects: link

Course project

- Apply the Bayesian methods you've learned in this course and/or explore new Bayesian methods to solve research question(s) of your choice
- Present at a poster session (last class), with a 2-min intro video
- Examples of projects from Spring 2019
  - Earnings and Sexual Orientation link
  - Bayes by Backprop: Weight Uncertainty in Neural Networks link
  - Bayesian Word Embeddings link
  - Full list of projects: link
- Questions?
- Ideas to discuss and share?

Grading

Homework & Labs	25%
Participation	10%
Midterm exam	40% (20% $\times$ 2)
Project	25%

- Grades curved at the end of the course after overall averages have been calculated.
- Average of 90-100 guaranteed A-.
- Average of 80-90 guaranteed B-.
- Average of 70-80 guaranteed C-.
- Average of 60-70 guaranteed D-.
- The more evidence there is that the class has mastered the material, the more generous the curve will be.

To do

- Complete the Class Survey and Pre-course Survey asap
- Register DataCamp account and access our MATH 347 Group
  - (Introduction to R - 4 hours)
  - (Intermediate R - 6 hours)
  - Introduction to the Tidyverse - 4 hours
  - All due Monday 9/16 11:59pm
- Make a self-introduction post (written or video) and read/watch others' posts (especially for common project interests - you can make a reply to indicate your interests of working on the project in pair)
- Get the textbook, read Chapter 1 and Chapter 2
- Download and install R and RStudio

Outline

Motivating Bayesian Inference

- Significance tests and confidence intervals are forms of classical or frequentist inference.
- When might classical inference be inadequate?
  - Suppose you flip a coin (with unknown probability of heads) three times and get tails all three times.
  - The sample percentage of heads equals zero. But, this can't be an accurate estimate of the true percentage of heads!
  - A priori of flipping the coin, we believe the true percentage is around 0.5, not 0.0.

Motivating Bayesian Inference

- Significance tests and confidence intervals are forms of classical or frequentist inference.
- When might classical inference be inadequate?
  - Suppose you flip a coin (with unknown probability of heads) three times and get tails all three times.
  - The sample percentage of heads equals zero. But, this can't be an accurate estimate of the true percentage of heads!
  - A priori of flipping the coin, we believe the true percentage is around 0.5, not 0.0.
- Bayesian inference provides a formal method for quantifying and incorporating our prior beliefs into inference.

Why Bayesian statistics?

- Long history: named after the 18^{th} century Presbyterian minister and mathematician Thomas Bayes (1701 - 1761).

Why Bayesian statistics?

- Long history: named after the 18^{th} century Presbyterian minister and mathematician Thomas Bayes (1701 - 1761).

- Modeling: incorporate prior belief or domain experts knowledge.
- Theoretical: doesn't need large sample assumption.
- Computational: Markov chain Monte Carlo (MCMC).

Bayesian approaches are largely popularized by revolutionary advance in computational technology during the last twenty five years (the invention of Gibbs sampler around 1990 - we will read a research paper about it later).

Two schools of statistics

Frequentist/classical

Probability: long run relative frequencies of repeatable events.

$$P(A) = \lim_{n\to\infty} \frac{\#(A)}{n}$$

- One time events?
- Small sample sizes?

Bayesian

Probability: a subjective degree of belief.

- Two people could have differing probabilities $P(A)$.
- Probability changes as new information (data) arise according to Bayes' rule.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

The classical thinking of probability is building on the assumption that you're able to repeat the experiments for a large number of times.

Two schools of statistics cont'd

Example 1: A study on several different, but similar populations, for each of which is needed an estimate of variability. How should one use the important prior information that the populations are similar?

Two schools of statistics cont'd

Example 1: A study on several different, but similar populations, for each of which is needed an estimate of variability. How should one use the important prior information that the populations are similar?

- Frequentist/classical: deciding between separate estimates of variance or a pooled variance $\rightarrow$ very crude utilization of the prior information
- Bayesian: techniques such as hierarchical Bayesian analysis to use the same distribution for all variances $\rightarrow$ allows much more efficient use of such prior information

Two schools of statistics cont'd

Example 2: A 95% confidence interval of an unknown parameter θ, interpretation? An interval that has probability of 0.95 of containing θ?

Two schools of statistics cont'd

Example 2: A 95% confidence interval of an unknown parameter θ, interpretation? An interval that has probability of 0.95 of containing θ?

Frequentist/classical: NO $\rightarrow$ indirectly related to the probability of the hypothesis

Bayesian: YES $\rightarrow$ directly related to the probability of the hypothesis

Classical: We are 95% CONFIDENT that the unknown parameter will fall into the interval. If we are able to perform the experiment for a large number of times, then about 95% intervals will cover the truth

Bayesian: We don't have confidence interval anymore. Instead, we have credible interval. A 95% credible interval of theta means that we have a 95% posterior probability that this interval will contain the theta.

Two schools of statistics cont'd

Two issues in the previous Example 2:

- Philosophically pragmatic: what is the best way of quantifying uncertainty?
- Pragmatically pragmatic: how statistical users (laymen) interpret statistical conclusions? How much work does it take for statistics educators to explain the correct interpretations of p-values and confidence intervals?

Two schools of statistics cont'd

Remarks:

- Not everyone has taken a mathematical statistics or classical statistical inference course, therefore in this course we won't emphasize more on the differences and comparisons between frequentist/classcial statistics and Bayesian statistics.
- We will think like a Bayesian for the entire semester.
- For those of you who had frequentist statistics training before, please think broadly and feel free to ask questions and/or start a discussion with the entire class (on Moodle)!

Toy example: procedure of Bayesian inference

Goal: making inference for an unknown quantity

Toy example: procedure of Bayesian inference

Goal: making inference for an unknown quantity

Inference procedure:

- Before seeing any data, we have prior probability: P(hypothesis)

Toy example: procedure of Bayesian inference

Goal: making inference for an unknown quantity

Inference procedure:

- Before seeing any data, we have prior probability: P(hypothesis)
- Collect data

Toy example: procedure of Bayesian inference

Goal: making inference for an unknown quantity

Inference procedure:

- Before seeing any data, we have prior probability: P(hypothesis)
- Collect data
- After collecting data, we update our probability of the hypothesis given the data we just observed. It is called posterior probability: P(hypothesis | data)

Toy example: procedure of Bayesian inference

Goal: making inference for an unknown quantity

Inference procedure:

- Before seeing any data, we have prior probability: P(hypothesis)
- Collect data
- After collecting data, we update our probability of the hypothesis given the data we just observed. It is called posterior probability: P(hypothesis | data)

Note that the frequentist p-value is P(data | hypothesis), the probability of observed or more extreme data given the null hypothesis being true.

Toy example: breast cancer screening

marginal probability

- American Cancer Society estimates that about 1.7% of women have breast cancer.
  http://www.cancer.org/cancer/cancerbasics/cancer-prevalence
- Susan G. Komen For The Cure Foundation states that mammography correctly identifies about 78% of women who truly have breast cancer.
  http://ww5.komen.org/BreastCancer/AccuracyofMammograms.html

conditional probability: true positive

- An article published in 2003 suggests that up to 10% of all mammograms are false positive. *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1360940*

偽陽性：If somebody does not have breast cancer , it will still come up positive as 10 percent

conditional probability: false positive

Note: *These percentages are approximate, and very difficult to estimate.*

Toy example: determining the prior

Prior to any testing and any information exchange between the patient and the doctor, what probability should a doctor assign to a female patient having breast cancer?

Toy example: determining the prior

Prior to any testing and any information exchange between the patient and the doctor, what probability should a doctor assign to a female patient having breast cancer?

0.017

Toy example: calculating the posterior

When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient has cancer, i.e. what is the posterior probability of having cancer if mammogram yield a positive result?

Toy example: calculating the posterior

When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient has cancer, i.e. what is the posterior probability of having cancer if mammogram yield a positive result?

P(true I positive)

$$P(c|+) = \frac{P(C \text{ and } +)}{P(+)}$$
$$= \frac{0.017 \times 0.78}{0.017 \times 0.78 + 0.983 \times 0.1}$$
$$= 0.12$$

P(+)=P(+ & Cancer)+P(+ & No Cancer)

Toy example: updating the prior

- In the Bayesian approach, we evaluate claims iteratively as we collect more data $\rightarrow$ sequential update.

Toy example: updating the prior

- In the Bayesian approach, we evaluate claims iteratively as we collect more data $\rightarrow$ sequential update.
- In the next iteration (screening) we get to take advantage of what we learned from the data.

Toy example: updating the prior

- In the Bayesian approach, we evaluate claims iteratively as we collect more data $\rightarrow$ sequential update.
- In the next iteration (screening) we get to take advantage of what we learned from the data.
- In other words, we update our prior with our posterior probability from the previous iteration. New prior would just be the previous posterior

Toy example: updating the prior when retesting

Suppose this woman who got a positive result in the first test wants to get tested again. What should the new prior probability that this woman has cancer? Is this probability smaller, larger, or equal to the prior probability in the first test? Why?

(a) 0.017

(b) 0.12

(c) 0.0133

(d) 0.88

Toy example: updating the prior when retesting

Suppose this woman who got a positive result in the first test wants to get tested again. What should the new prior probability that this woman has cancer? Is this probability smaller, larger, or equal to the prior probability in the first test? Why?

(a) 0.017

(b) 0.12

(c) 0.0133

(d) 0.88

Answer: (b)

Toy example: re-calculating the posterior when retesting

What is the posterior probability of having cancer if this second mammogram also yielded a positive result?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52

Toy example: re-calculating the posterior when retesting

What is the posterior probability of having cancer if this second mammogram also yielded a positive result?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52

Answer: (d)

$$P(c|+) = \frac{P(c \text{ and } +)}{P(+)} = \frac{0.12 \times 0.78}{0.12 \times 0.78 + 0.88 \times 0.1} = 0.52$$

Toy example: recap of Bayesian inference

- Take advantage of prior information, like a previously published study or a physical model.

Toy example: recap of Bayesian inference

- Take advantage of prior information, like a previously published study or a physical model.
- Naturally integrate data as you collect it, and update your priors.

Toy example: recap of Bayesian inference

- Take advantage of prior information, like a previously published study or a physical model.
- Naturally integrate data as you collect it, and update your priors.
- Avoid the counter-intuitive Frequentist definition of a p-value as the P(observed or more extreme outcome | H_0 is true). Instead base decisions on the posterior probability, P(hypothesis is true | observed data).

Toy example: recap of Bayesian inference

- Take advantage of prior information, like a previously published study or a physical model.
- Naturally integrate data as you collect it, and update your priors.
- Avoid the counter-intuitive Frequentist definition of a p-value as the P(observed or more extreme outcome | H_0 is true). Instead base decisions on the posterior probability, P(hypothesis is true | observed data).
- Watch out! A good prior helps, a bad prior hurts, but the prior matters less the more data you have.

Toy example: recap of Bayesian inference

- Take advantage of prior information, like a previously published study or a physical model.
- Naturally integrate data as you collect it, and update your priors.
- Avoid the counter-intuitive Frequentist definition of a p-value as the P(observed or more extreme outcome | H_0 is true). Instead base decisions on the posterior probability, P(hypothesis is true | observed data).
- Watch out! A good prior helps, a bad prior hurts, but the prior matters less the more data you have.
- More advanced Bayesian techniques offer flexibility not present in Frequentist models.

Outline

What is fixed, Y or θ?

Both frequentist and Bayesian aim to use data Y to learn about the unknown parameter θ.

Frequentist:

$$p(Y \mid \theta)$$

- Data are a repeatable random sample
- Underlying parameters remain constant during this repeatable process
- Parameters θ are fixed

Bayesian:

$$p(\theta \mid Y)$$

- Data are observed from the realized sample
- Parameters are unknown and described probabilistically
- Data Y are fixed

Parameters are random

Extending to statistical analysis

Bayes Theorem extends naturally to parameters in statistical inference.

- "Characteristics" are akin to parameters θ in probability models, e.g., $\theta = p$ in the Binomial distribution $\text{Binomial}(n, p)$.
- "Data" are akin to measurements on sampled data subjects expressed numerically, say y.
- Before the sample is collected, both y and θ are unknown.
- "Model for data" is akin to a probability model for Y assuming we know θ, e.g., $Y \sim \text{Binomial}(n, p)$.
- We express prior information about θ as a (different) probability model.

Bayesian inference

Bayesian inference provides a formal approach for updating prior beliefs with the observed data to quantify uncertainty a posteriori about θ.

- Prior distribution $\pi(\theta)$
- Sampling model $f(y \mid \theta)$
- Posterior distribution:

$$\pi(\theta \mid y) = \frac{f(y \mid \theta)\,\pi(\theta)}{f(y)} = \frac{f(y \mid \theta)\,\pi(\theta)}{\int_\Theta f(Y \mid \tilde{\theta})\pi(\tilde{\theta})d\tilde{\theta}}$$

f(y | \tilde \theta) \pi(\tilde \theta) = f(y, \tilde \theta)

(for discrete support for θ replace integral with sum)

Bayesian inference for proportion

- What percentage p of all Vassar students stayed up at least one night to get school work done last academic year.
- Suppose we sample ten MATH 347 students and ask whether they stayed up at least one night last academic year.

Bayesian inference for proportion

- What percentage p of all Vassar students stayed up at least one night to get school work done last academic year.
- Suppose we sample ten MATH 347 students and ask whether they stayed up at least one night last academic year.
- Let's build a (simplified) prior distribution and use Bayesian inference to learn about $\theta = p$.
- We will make the (incorrect) assumption that the sample is representative of Vassar students at large, since this is just an illustration of ideas.

Prior distribution

We will use the unrealistic but instructive prior distribution defined by consensus of the class:

$$
\begin{aligned}
P(p = 0) = \quad & P(p = 0.1) = \quad & P(p = 0.2) = \\
P(p = 0.3) = \quad & P(p = 0.4) = \quad & P(p = 0.5) = \\
P(p = 0.6) = \quad & P(p = 0.7) = \quad & P(p = 0.8) = \\
P(p = 0.9) = \quad & P(p = 1.0) = &
\end{aligned}
$$

We'll do the posterior probability calculation following the All_nighters_example.Rmd (available on Moodle).

Analsyis goals

Bayesian methods go beyond the formal updating of the prior distribution to obtain a posterior distribution:

- Estimation of uncertain quantities (parameters) with good statistical properties.
- Prediction of future events.
- Tests of hypotheses.
- Making decisions.

Outline

Events and partitions

Definition

A collection of sets $\{H_1, \cdots, H_k\}$ *is a partition of another set* $\mathcal{H}$ *if*

1. *the events are disjoint, which we write as* $H_i \cap H_j = \emptyset$ *for* $i \neq j$*;*
2. *the union of the sets is* $\mathcal{H}$*, which we write as* $\cup_{k=1}^{K} H_k = \mathcal{H}$*.*

Examples?

Partitions and probability

Suppose $\{H_1, \cdots, H_K\}$ is a partition of $\mathcal{H}$, $Pr(\mathcal{H}) = 1$, and E is some specific event. The axioms of probability imply:

- Rule of total probability
$$\sum_{k=1}^{K} Pr(H_k) = 1$$
- Rule of marginal probability
$$Pr(E) = \sum_{k=1}^{K} Pr(E \cap H_k) = \sum_{k=1}^{K} Pr(E \mid H_k) Pr(H_k)$$
- Bayes' rule
$$Pr(H_j|E) = \frac{Pr(E \mid H_j) Pr(H_j)}{Pr(E)} = \frac{Pr(E \mid H_j) Pr(H_j)}{\sum_{k=1}^{K} Pr(E \mid H_k) Pr(H_k)}$$

In Bayesian inference

- $\{H_1, \cdots, H_K\}$: disjoint hypotheses or states of nature.
- E: the outcome of a survey, study or experiment.

To compare hypotheses post-experimentally $Pr(H_i \mid E)$ v.s. $Pr(H_j \mid E)$:

$$\frac{Pr(H_i \mid E)}{Pr(H_j \mid E)} = \frac{Pr(E \mid H_i)Pr(H_i)/Pr(E)}{Pr(E \mid H_j)Pr(H_j)/Pr(E)}$$

In Bayesian inference

- $\{H_1, \cdots, H_K\}$: disjoint hypotheses or states of nature.
- E: the outcome of a survey, study or experiment.

To compare hypotheses post-experimentally $Pr(H_i \mid E)$ v.s. $Pr(H_j \mid E)$:

$$\begin{aligned} \frac{Pr(H_i \mid E)}{Pr(H_j \mid E)} &= \frac{Pr(E \mid H_i)Pr(H_i)/Pr(E)}{Pr(E \mid H_j)Pr(H_j)/Pr(E)} \\ &= \frac{Pr(E \mid H_i)Pr(H_i)}{Pr(E \mid H_j)Pr(H_j)} \end{aligned}$$

In Bayesian inference

- $\{H_1, \cdots, H_K\}$: disjoint hypotheses or states of nature.
- E: the outcome of a survey, study or experiment.

To compare hypotheses post-experimentally $Pr(H_i \mid E)$ v.s. $Pr(H_j \mid E)$:

$$\begin{aligned}
\frac{Pr(H_i \mid E)}{Pr(H_j \mid E)} &= \frac{Pr(E \mid H_i)Pr(H_i)/Pr(E)}{Pr(E \mid H_j)Pr(H_j)/Pr(E)} \\
&= \frac{Pr(E \mid H_i)Pr(H_i)}{Pr(E \mid H_j)Pr(H_j)} \\
&= \frac{Pr(E \mid H_i)}{Pr(E \mid H_j)} \times \frac{Pr(H_i)}{Pr(H_j)} \\
&= \text{"}\textit{Bayes factor}\text{"} \times \text{"}\textit{prior beliefs}\text{"}
\end{aligned}$$

In Bayesian inference

Think of H as hypothesis

- $\{H_1, \cdots, H_K\}$: disjoint hypotheses or states of nature.
- E: the outcome of a survey, study or experiment. Think of E as data

To compare hypotheses post-experimentally $Pr(H_i \mid E)$ v.s. $Pr(H_j \mid E)$:

$$\begin{aligned}
\frac{Pr(H_i \mid E)}{Pr(H_j \mid E)} &= \frac{Pr(E \mid H_i)Pr(H_i)/Pr(E)}{Pr(E \mid H_j)Pr(H_j)/Pr(E)} \\
&= \frac{Pr(E \mid H_i)Pr(H_i)}{Pr(E \mid H_j)Pr(H_j)} \\
&= \frac{Pr(E \mid H_i)}{Pr(E \mid H_j)} \times \frac{Pr(H_i)}{Pr(H_j)} \\
&= \text{"}\textit{Bayes factor}\text{"} \times \text{"}\textit{prior beliefs}\text{"}
\end{aligned}$$

LHS= ratio of posterior

Bayes' rule does not determine what our beliefs should be after seeing the data. It tells us how they should change after seeing the data.

Bayes factor help you to change/update your belief

Univariate random variable

Cumulative distribution function (cdf)

$$F_X(x) = P(X \leq x), \quad \text{for any } x \in \mathbb{R}$$

	Discrete	Continuous
pmf / pdf	$f_X(x) = P(X = x)$	$P(X \in B) = \int_B f(x)dx$
well-defined	$\sum_{i=1}^{\infty} f(x_i) = 1$	$\int_{-\infty}^{\infty} f(x)\ dx = 1$
cdf	$F(a) = \sum_{\text{all } x \leq a} f(x)$	$F(x) = \int_{-\infty}^{x} f(t)dt$ $f(x) = \frac{d}{dx}F(x)$
expectation	$E[X] = \sum_{\text{all } x} x \cdot f(x)$	$E[X] = \int_{-\infty}^{\infty} xf(x)dx$
	$E[g(X)] = \sum_{\text{all } x} g(x)\ f(x)$	$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$

Joint distribution

	Discrete	Continuous
pmf/pdf	$P(X = x, Y = y)$	$P[(X, Y) \in C] = \iint_{(x,y)\in C} f(x, y)\ dxdy$
marginal	$f_X(x) = \sum_y f(x, y)$,	$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\ dy$,
	$f_Y(y) = \sum_x f(x, y)$	$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\ dx$
cdf		$F(a, b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x, y)\ dxdy$
		$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$

Conditional distribution

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}$$

- Joint density $f(x,y) = f(x \mid y) f_Y(y) = f(y \mid x) f_X(x)$

Conditional distribution

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}$$

- Joint density $f(x,y) = f(x \mid y)f_Y(y) = f(y \mid x)f_X(x)$

Conditional expectation

- $E(X \mid Y = y) = \sum_x f(x \mid y)$ or $\int_{-\infty}^{\infty} xf(x \mid y)dx$
- Law of total expectation

$$E_Y\left[E(X \mid Y)\right] = E(X)$$

Conditional variance

- $Var(X \mid Y = y) = E[(X - E[X \mid Y = y])^2 \mid Y = y]$
- A very useful conditional variance formula

$$Var(X) = E[Var(X \mid Y)] + Var(E[X \mid Y])$$

Transformation of variables

Suppose X is a continuous random variable with pdf $f_X(x)$. If a function $g(x)$ is

- monotonic (increasing or decreasing), and
- differentiable (and thus continuous),

then the random variable defined by $Y = g(X)$ has pdf

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$$

Transformation of variables

Suppose X is a continuous random variable with pdf $f_X(x)$. If a function $g(x)$ is

- monotonic (increasing or decreasing), and
- differentiable (and thus continuous),

x=g^{-1}(y)

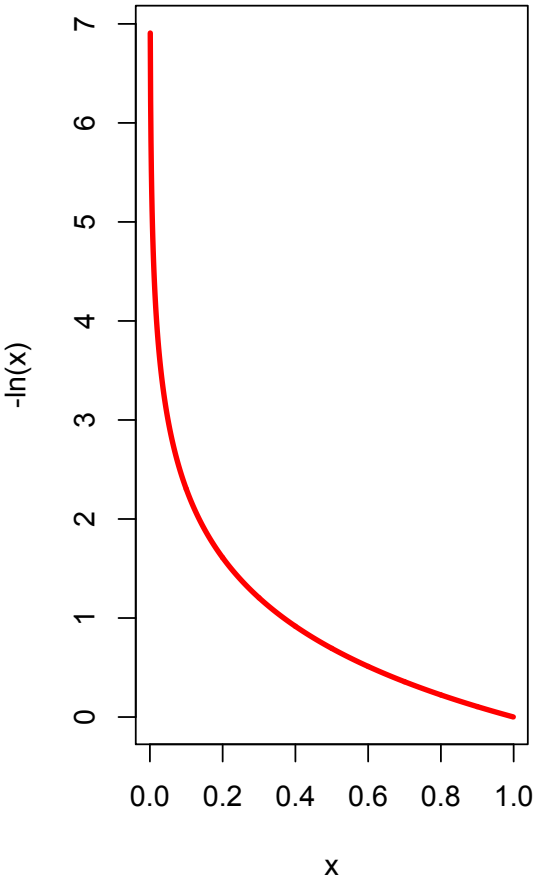
then the random variable defined by $Y = g(X)$ has pdf

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

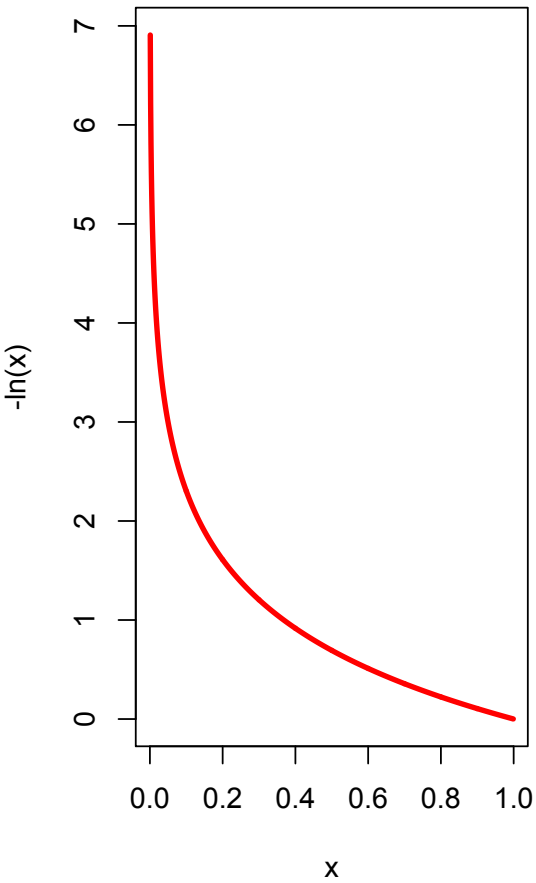
Or more rigorously,

$$f_Y(y) = \begin{cases} f_X\left[g^{-1}(y)\right] \cdot \left|\frac{d}{dy} g^{-1}(y)\right| & \text{if } y = g(x) \text{ for some } x \\ 0 & \text{if } y \neq g(x) \text{ for all } x \end{cases}$$

Example: let $X \sim \text{Unif}(0, 1)$, what distribution does $Y = -\ln(X)$ have?
(Hint: monotonic & differentiable, then $f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$)



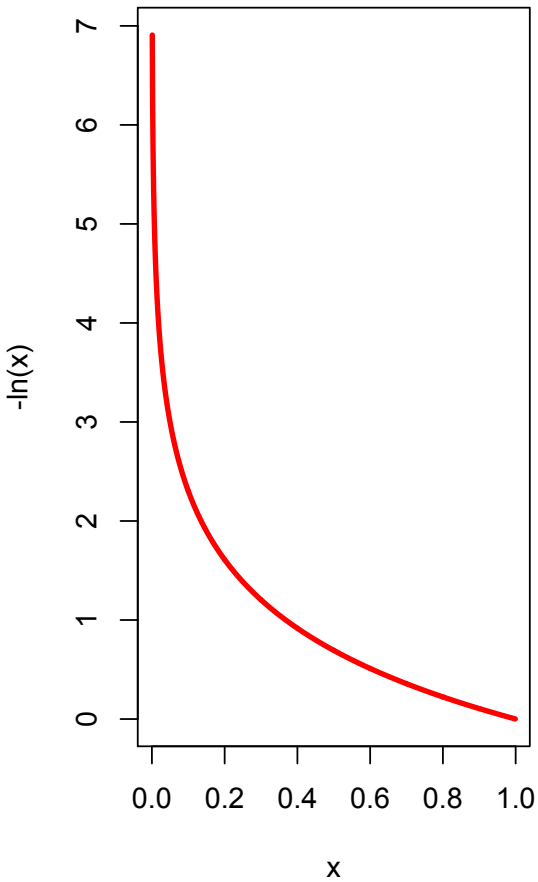
Example: let $X \sim \text{Unif}(0,1)$, what distribution does $Y = -\ln(X)$ have? (Hint: monotonic & differentiable, then $f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$)



In order to use the previous theorem, need to check the function $g(x) = -\ln(x)$

- monotonic
- differentiable

Example: let $X \sim \text{Unif}(0,1)$, what distribution does $Y = -\ln(X)$ have? (Hint: monotonic & differentiable, then $f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$)



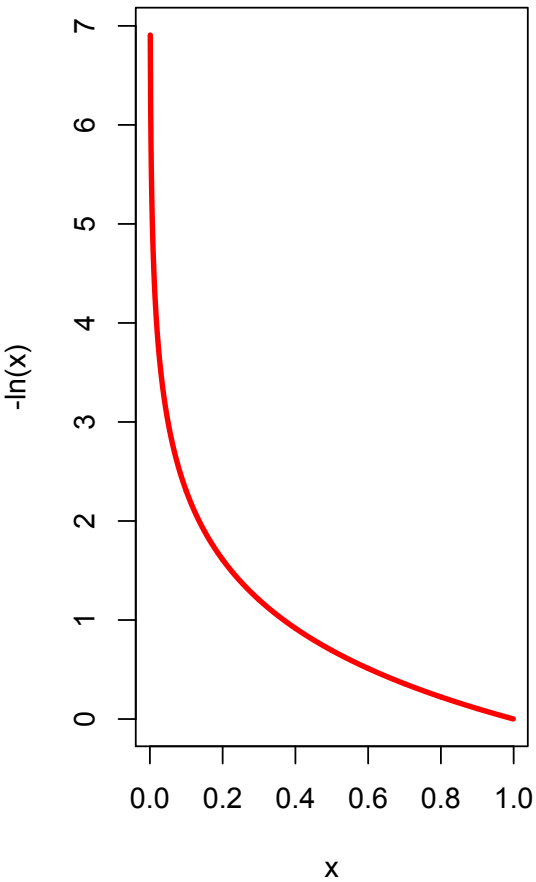
In order to use the previous theorem, need to check the function $g(x) = -\ln(x)$

- monotonic
- differentiable

Inverse function $g^{-1}(y)$: $y = -\ln(x) \iff x = e^{-y}$

$$\frac{dx}{dy} = -e^{-y}$$

Example: let $X \sim \text{Unif}(0,1)$, what distribution does $Y = -\ln(X)$ have? (Hint: monotonic & differentiable, then $f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$)



In order to use the previous theorem, need to check the function $g(x) = -\ln(x)$

- monotonic
- differentiable

Inverse function $g^{-1}(y)$: $y = -\ln(x) \iff x = e^{-y}$

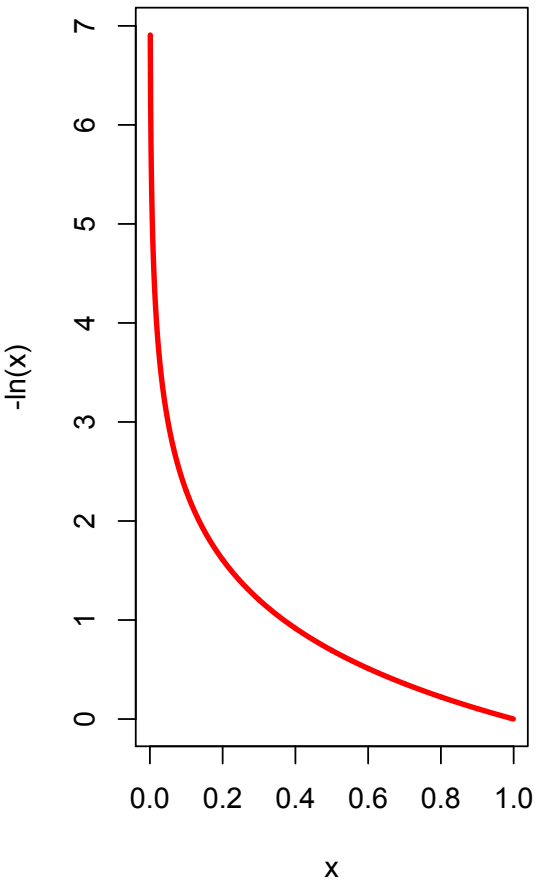
$$\frac{dx}{dy} = -e^{-y}$$

Range of Y: $y \in [0, \infty)$.

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = 1 \cdot \left|-e^{-y}\right| = e^{-y}$$

f(x)=1, if 0< x <1, =0, o.w.

Example: let $X \sim \text{Unif}(0,1)$, what distribution does $Y = -\ln(X)$ have? (Hint: monotonic & differentiable, then $f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$)



In order to use the previous theorem, need to check the function $g(x) = -\ln(x)$

- monotonic
- differentiable

Inverse function $g^{-1}(y)$: $y = -\ln(x) \iff x = e^{-y}$

$$\frac{dx}{dy} = -e^{-y}$$

Range of Y: $y \in [0, \infty)$. 0 < x < 1 implies -Inf < ln(x) < 0

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = 1 \cdot \left|-e^{-y}\right| = e^{-y}$$

Therefor, Y has an exponential distribution: $Y \sim \text{Exp}(1)$.

Independent random variables

Definition

Suppose $Y_1, \cdots, Y_n$ are random variables and that θ is a parameter describing the conditions under which the random variables are generated. We say that $Y_1, \cdots, Y_n$ are conditionally independent given θ if for every collection of n sets $\{A_1, \cdots, A_n\}$ we have

$$Pr(Y_1 \in A_1, \cdots, Y_n \in A_n \mid \theta) = Pr(Y_1 \in A_1 \mid \theta) \times \cdots \times Pr(Y_n \in A_n \mid \theta)$$

Independent random variables

Definition

Suppose $Y_1, \cdots, Y_n$ are random variables and that θ is a parameter describing the conditions under which the random variables are generated. We say that $Y_1, \cdots, Y_n$ are conditionally independent given θ if for every collection of n sets $\{A_1, \cdots, A_n\}$ we have

$$Pr(Y_1 \in A_1, \cdots, Y_n \in A_n \mid \theta) = Pr(Y_1 \in A_1 \mid \theta) \times \cdots \times Pr(Y_n \in A_n \mid \theta)$$

Under independence, the joint density is the product of the marginal densities

$$f(y_1, \cdots, y_n \mid \theta) = f_{Y_1}(y_1 \mid \theta) \times \cdots \times f_{Y_n}(y_n \mid \theta) = \prod_{i=1}^{n} f_{Y_i}(y_i \mid \theta)$$

Independent random variables

Definition

Suppose $Y_1, \cdots, Y_n$ are random variables and that θ is a parameter describing the conditions under which the random variables are generated. We say that $Y_1, \cdots, Y_n$ are conditionally independent given θ if for every collection of n sets $\{A_1, \cdots, A_n\}$ we have

$$Pr(Y_1 \in A_1, \cdots, Y_n \in A_n \mid \theta) = Pr(Y_1 \in A_1 \mid \theta) \times \cdots \times Pr(Y_n \in A_n \mid \theta)$$

Under independence, the joint density is the product of the marginal densities

$$f(y_1, \cdots, y_n \mid \theta) = f_{Y_1}(y_1 \mid \theta) \times \cdots \times f_{Y_n}(y_n \mid \theta) = \prod_{i=1}^{n} f_{Y_i}(y_i \mid \theta)$$

Definition

In this case, we say that $Y_1, \cdots, Y_n$ are conditionally independent and identically distributed (i.i.d.)

$$Y_1, \cdots, Y_n \mid \theta \overset{iid}{\sim} f(y \mid \theta)$$

Independent random variables cont'd

Example: let $Y_1, \cdots, Y_n \mid \mu, \sigma \overset{iid}{\sim} \text{Normal}(\mu, \sigma)$, write out the joint density of $f(Y_1, \cdots, Y_n \mid \mu, \sigma)$. (Hint: the Normal pdf is $f(Y_i = y_i \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i-\mu)^2}{2\sigma^2}))$

Independent random variables cont'd

Example: let $Y_1, \cdots, Y_n \mid \mu, \sigma \overset{iid}{\sim} \text{Normal}(\mu, \sigma)$, write out the joint density of $f(Y_1, \cdots, Y_n \mid \mu, \sigma)$. (Hint: the Normal pdf is $f(Y_i = y_i \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i-\mu)^2}{2\sigma^2}))$

$$
\begin{aligned}
f(Y_1 = y_1, \cdots, Y_n = y_n \mid \mu, \sigma) &= \prod_{i=1}^{n} f_{Y_i}(y_i \mid \mu, \sigma) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum(y_i-\mu)^2}{2\sigma^2}\right)
\end{aligned}
$$

Exchangeablility

Let $f(y_1, \ldots, y_n)$ be the joint distribution of $Y_1, \ldots, Y_n$ and let $\pi_1, \ldots, \pi_n$ be a permutation of the indices $1, \ldots, n$. If $f(y_1, \ldots, y_n) = f(y_{\pi_1}, \ldots, y_{\pi_n})$ for all permutations, then $Y_1, \ldots, Y_n$ are exchangeable.

de Finetti's Theorem

$Y_1, \cdots, Y_n$ be a sequence of random variables. If for any n, $Y_1, \cdots, Y_n$ are exchangeable, then there exists a prior distribution $\pi(\theta)$ and sampling model $f(y \mid \theta)$ such that

$$f(y_1, \ldots, y_n) = \int_\Theta \left\{ \prod_1^n f(y_i \mid \theta) \right\} \pi(\theta)\, d\theta$$

Models

$$\left.\begin{array}{l} Y_1, \ldots Y_n \mid \theta \overset{\text{iid}}{\sim} p(y \mid \theta) \\ \theta \sim p(\theta) \end{array}\right\} \iff Y_1, \ldots Y_n \text{ are exchangeable for all } n$$

Applicable if $Y_1, \ldots, Y_n$ are

- outcomes of a repeatable experiment
- random sample from finite population with replacement
- sampled from an infinite population w/out replacement
- sampled from a finite population of size $N >> n$ w/out replacement (approximate)

Labels carry no information.