Journal of Statistics Education

Large Deviations: Advanced Probability for Undergrads

David A. Rolls
The University of Melbourne, Australia

Journal of Statistics Education Volume 15, Number 3 (2007),
www.amstat.org/publications/jse/v15n3/rolls.html

Key Words: convergence, Chernoff's Theorem, Cramer's Theorem, independent study, teaching, undergraduate statistics

Abstract

In the branch of probability called "large deviations," rates of convergence (e.g. of the sample mean) are considered. The theory makes use of the moment generating function. So, particularly for sums of independent and identically distributed random variables, the theory can be made accessible to senior undergraduates after a first course in stochastic processes. This paper describes a directed independent study in large deviations offered to a strong senior, providing a sample outline and discussion of resources. Learning points are also highlighted.

Introduction

Imagine flipping a fair coin many times. We model the result of flip n with random variable X_n where `0' and `1' denote a head and a tail, respectively. For a fair coin $P(X_n=0) = P(X_n=1) = 0.5$. We model the collection of flips with the sequence $\{X_n, n = 1,2,...\}$ which is an independent and identically distributed (i.i.d.) sequence with common mean $E[X_n] = 0.5$. Let $S_n = \sum^n_{i=1} X_i$, so S_n counts the number of tails in n flips. By the Strong Law of Large Numbers, with probability 1, $S_n/n \to 0.5$ as $n \to \infty$. In other words, if the number of flips is large the proportion of tails will be about half.

For any $n \geq 1$, S_n/n is a random variable too. Since $S_n/n \to 0.5$ as $n \to \infty$, for any $a > 0.5$,

$$P(S_n \geq na) \to 0 \text{ as } n \to \infty.$$

An interesting question is *how quickly* does $P(S_n \geq na)$ go to 0? For large n, S_n/n significantly different from 0.5 would be considered a "rare event". But what is "significant"? These and similar questions are answered in the branch of probability called "large deviations," with applications to areas such as

statistics, connection admission control in data networks, and the study of polymers.

While usually taught at the graduate level, aspects of the theory can be made accessible to senior undergrads, particularly when the sequence $\{X_n, n = 1,2,...\}$ is i.i.d. The main prerequisite is previous exposure to the moment generating function. This paper follows my supervision of a senior undergrad, Rick (not his real name) in an independent study at the University of North Carolina Wilmington who asked specifically for exposure to more sophisticated ideas of probability as he was bound for a graduate statistics program. The remainder of this article gives an overview of large deviations and available resources appropriate for senior undergrads with some previous exposure to real analysis (e.g., limsup, liminf, compact sets). Large deviations can become quite technical. (The books by den Hollander (2000) and Dembo and Zeitouni (1998) are excellent references.) The goal here is not to present those details, but rather to show a route through some basic ideas for undergrads that avoids some of them, while presenting sufficient challenges.

The rest of this paper is organized as follows. The remainder of Section 1 discusses a number of resources that are available in the area of large deviations. Section 2 provides more details and questions on the coin tossing example. Section 3 discusses Cramér's theorem, and describes some of the discussion points one could raise with an undergraduate. Section 4 discusses the project and how it progressed. Section 5 gives some concluding remarks, while Section 6 provides the outline for the project.

1.1 Resources

There are a number of references for large deviations and the quality and level of the exposition varies. All four of the books listed here are texts, and include questions which can be assigned. Careful choices must be made to keep the difficulty appropriate.

A general introduction to the topic for discrete random variables can be found in Section 9 of Billingsley (1986). Generally the level of that book is at a more advanced level but this optional section is something of an anomaly. Starting from the moment generating function it builds up to a version of an important theorem, Chernoff's theorem, and gives an example in the context of statistical hypothesis testing.

The text by den Hollander (2000) is split roughly 50-50 between theory and applications. Chapter 1 (8 pages) provides introductory theory for i.i.d. sequences. It is rigorous without excessive detail. It does assume students know some basics about measures. For example, if X is a random variable with cumulative distribution function F, so that $F(x) = P(X \leq x)$, students are assumed to understand integrals like $\int_R g(x)dF(x)$. This need not be a barrier to undergrads however. To make this notation accessible there are a couple of possibilities. They can be told this is simply $E[g(x)]$. Or, you can stipulate an assumption that X is (absolutely) continuous with density function f so that $\int_R g(x)dF(x) = \int_R g(x)f(x)dx$. In this form the expected value is recognizable. Chapter 6 (4 pages) provides an accessible application to statistical hypothesis testing with a nice statement of the statistical problem of interest and a quick overview of Neyman-Pearson tests. Chernoff's theorem is used to find the exponential rate of decay of error probabilities with increasing sample size n in optimal Neyman-Pearson tests.

The text by Bucklew (1990) is written at an upper graduate student level with applications to information theory, communications theory or applied statistics. For undergraduates, the introduction (Bucklew 1990, pp. 1-4) is accessible and provides motivation. Two examples on Chernoff's theorem (Bucklew 1990, p. 35) are also accessible.

The text by Dembo & Zeitouni (1998) is a staple in large deviations. It is mathematical, thorough and goes well beyond the i.i.d. setting. So, most of the coverage is too advanced for use here. Even the notation could prove challenging. One section of the text, Section 2.2.1, deals with Cramér's Theorem in

the i.i.d. setting. If a student is given a more accessible presentation beforehand, this section provides a glimpse into a more sophisticated presentation of the material. It is expected they would have to read and re-read the section, asking questions for clarification. To help decipher some notation, I told Rick that for a set A

$$\mu_n(A) \equiv P(\hat{S}_n \in A)$$

where

$$\hat{S}_n$$

is the empirical mean in that section.

Finally, I included two papers in the outline of the project. The paper by Weiss (1995) is a very accessible introduction, written as an introduction for engineers. The paper by Chernoff (1952) illustrates the use of large deviations to the analysis of statistical hypothesis testing. In the project, reading Chernoff's paper is a choice. In fact, the texts above provide more accessible treatments to this application. While Rick was initially keen to read it, when it was time to choose later in the semester, he went with the text treatment.

2. Coin Tossing

The coin tossing example above leads to the following

Lemma 2.1 (See den Hollander (2000)). *Let* $\{X_n, n = 1,2,...\}$ *be an i.i.d. sequence with* $P(X_n = 0) = P(X_n = 1) = 0.5$. *Then for all* $a > 0.5$

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n \geq na) = -I(a)$$

where

$$I(x) = \begin{cases} \log 2 + x \log x + (1-x)\log(1-x) & 0 \leq x \leq 1 \\ \infty & \text{otherwise.} \end{cases}$$

A direct proof of this result is possible and makes a good exercise. The interesting case has $0.5 < a \leq 1$, uses a counting argument, Stirling's approximation

$$n! \sim n^n e^{-n} \sqrt{2\pi n}$$

and the inequalities

$$2^{-n} Q_n(a) \leq P(S_n \geq na) \leq (n+1) 2^{-n} Q_n(a)$$

where

$$Q_n(a) = \max_{k \geq na} \binom{n}{k}.$$

(The maximum is attained at $k = \lceil na \rceil$. A proof appears in den Hollander (2000, p. 5), but even that proof is sufficiently terse that students are sufficiently challenged if asked to give full details. For example, how does one correctly use the asymptotic result from Stirling's approximation? (It is *not true* that

$$n! = n^n e^{-n} \sqrt{2\pi n}$$

- the result is asymptotic.) Why are the inequalities in (3) true? How does one formally work with the ceiling function in $\lceil na \rceil$ (e.g., for the asymptotics of $\lceil na \rceil$ as n '! ")?

3. Cramér's Theorem

The first fundamental result of large deviation theory is Cramér's Theorem, (also called Chernoff's Theorem) and makes use of the moment generating function $M(t)$ given by

$$M(t) = E[e^{tX}].$$

The Fenchel-Legendre transform is

$$I(x) = \sup_t \{t\,x - \log M(t)\}.$$

Cramér's Theorem. *Let* $\{X_n, n = 1,2,...\}$ *be an i.i.d. sequence with* $E[X_1] = \mu$ *and* $M(t) < \infty$ *for all* t.

For all a > μ

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n \geq na) = -I(a).$$

Proof. See den Hollander (2000, p. 5)

This theorem includes the coin tossing example, but is much more general. Calculating $I(x)$ for common distributions like Bernoulli(p), exponential(λ), Poisson(λ) and Normal(μ, σ^2) is a nice application of finding absolute maximums. Students, such as Rick, can be good at finding critical numbers, but may need reminding to justify they are absolute maximums.

Definition. *If* $y_1, y_2,...$ *is a sequence of numbers such that* $y_n \to y$ *as* $n \to \infty$ *, a function f is lower semi-continuous if* $\liminf_{n\to\infty} f(y_n) \geq f(y)$.

The function I has a number of properties, including 1) I is non-negative, 2) I is lower semi-continuous, 3) $I(x) = 0$ iff $x = E(X_1)$, and 4) I is convex. The level sets of I are the sets $I^{-1}([0,a]) = \{x: I(x) \leq a\}$. Lower semi-continuous is equivalent to the level sets of I being closed. A proof of property (1) is trivial. A proof for property (2) is a nice application of the following lemma.

Fatou's Lemma. *If $\{Y_n, n = 1,2,...\}$ is a sequence of non-negative valued random variables with finite mean*

$$E[\liminf_{n\to\infty} Y_n] \le \liminf_{n\to\infty} E[Y_n].$$

(Note that non-negativity in Fatou's lemma doesn't mean we can only consider non-negative sequences $\{X_n, n = 1,2,...\}$. The connection between X_n and Y_n is more subtle.) This also provides a nice opportunity to briefly introduce the limit theorems for integrals (monotone convergence, dominated convergence) and place Fatou's lemma in context. For a sequence $\{Z_n, n = 1,2,...\}$ of continuous random variables the statement $E[\lim_{n\to\infty} Z_n] = \lim_{n\to\infty} E[Z_n]$

isn't necessarily true. Similarly, for a sequence of functions $\{f_n,1,2,...\}$ where $f_n \to f$ as $n \to \infty$ it is not necessarily true that $\lim_{n\to\infty} \int f_n(x)\,dx = \int f(x)\,dx.$

These results are generally established with one of the "convergence theorems". These theorems have stronger requirements on the sequence than Fatou's Lemma. In this sense Fatou's Lemma is a lesser alternative- a weaker hypothesis, but providing a weaker result. It is used less often. But that weaker result is still useful to prove property (2)! For this first exposure, these convergence results need not be proven to the students. For a proof of property (3), Jensen's inequality is useful, which will probably require a hint.

By the properties of I, it is not decreasing on (μ, ∞). The result of Cramér's Theorem can be written (den Hollander 2000, p. 10) as

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in A\right) = -I(a) = -\inf_{x\in A} I(x) \text{ where } A \in [a, \infty).$$

It may be surprising to see that the probability is obtained from the minimizing value x^* in A which is closest to μ (i.e. the lower endpoint a). But this is the fundamental observation of large deviations: "rare events happen in the least unlikely of the unlikely ways".

4. Teaching Large Deviations

Section 6 provides a syllabus with timetable for the project. It shows there were really three "phases" to Rick's independent study. The times shown were the ones I originally scheduled for the project, and worked well until the written report, which needed two additional weeks, but allowed some flexibility as Rick's regular courses were winding down- some with final exams. The first phase of the project, Section 6.4, about 8 weeks long, includes a number of readings, calculations and proofs. I asked Rick to redo some solutions and proofs to achieve required precision (e.g., problems 1 and 4) which I think is a good learning tool. Generally, I gave feedback on his submissions within a few days, but up to a week was possible.

Through this phase Rick worked hard, perhaps harder than he was used to. Through the entire semester we had a regularly-scheduled time each week when he would come by to ask questions, report progress or submit solutions. But I was accessible at other times too, and he would use those times to ask additional questions. This first phase generated the most questions from him, and led to the most

face-to-face time. In some cases he asked my colleagues, whom he knew well, about supplemental texts for foundational probability or real analysis. He was challenged, but I think he rose to the occasion. Never did he appear discouraged.

The second phase, described in Section 6.5, was about 4 weeks long and involved computer simulation work. It was intentionally designed to be not too onerous since learning the necessary computer skills can take time. Rick chose to use R, with which he had some previous experience. Large deviation theory gives an exponential decay rate. The point here was to use simulations to demonstrate that rate. This required some mentoring, but far less involvement by me than in the other two phases. Since Rick had a lot of training in statistics he was able to give both point estimates and confidence intervals for that rate. Happily, all his confidence intervals covered the theoretical rate.

The final phase, Section 6.6, was the report. Rick's final report was 15 pages (double spaced) plus appendices for computer code and data. He chose to use Microsoft Word to prepare the document. The assignment is intentionally written to require some editorial decisions on what is worth including. It was intended that Rick describe what he learned in a careful mathematical way. This turned out to be a challenging phase, taking about three weeks and five drafts. To graduate Rick didn't need to write a senior thesis. So, some of my comments were either about English writing in general or typical comments for someone with little to no experience writing reports in mathematics. In some cases, I had to prod him to include more of the work he had done. In addition, some teaching points from the first phase were clearly not mastered, and the report was an opportunity to reinforce them. Stirling's approximation is not an equality. Chernoff's theorem is an asymptotic result, not an inequality for every n. There is a difference between the rate function for a zero threshold Neyman-Pearson test and the test itself. I expected some points would not be mastered the first time around. This is one reason I assigned a written report over an oral presentation- an oral presentation doesn't lend itself to revisiting points for mastery. So the learning continued, basically right through until the last week.

With hindsight, I am reasonably happy with how phase 1 of the project developed. Finding the right balance of giving hints versus having a student try again is challenging but important if the student is to stay motivated. I tend to think Rick could have handled a bit more frustration from trying again.

I can see a few things I would do differently. To me, the written report is a valuable part of the assignment. A nice advantage is that, as far as I know, cut-and-paste from web resources isn't really possible. But it realistically requires 3-4 weeks for students with little writing experience, even with quick feedback for each draft. That's to achieve a report that a student could proudly show to other people. Reduced expectations might eliminate some iterations of the revisions process. To provide the extra time, a week (maybe two) can be borrowed from the computer simulation time (phase 2) if a student is using software they have used before.

For the report, Word is cumbersome because of the volume and complexity of math content. LaTeX would be a natural choice, but the learning curve is too steep for student use with no previous experience. Next time, I would have a student try writing the report using "document mode" in the latest versions of Maple- assuming some prior familiarity with Maple. This allows presentation-quality documents to be prepared using a familiar Maple interface, syntax, symbol palettes and so-on.

The use of the Normal distribution in phase 2 was a deliberate choice, but ultimately it wasn't fully exploited. In fact, the distribution of S_n is known exactly and one can compare exact results for $P(S_n \geq na)$ versus n as $n \to \infty$ with large deviation asymptotic results and those obtained from the estimates. Alternatively, one can use a non-Gaussian distribution where exact results are not known.

5. Conclusion

An independent study course in large deviations is not appropriate for every undergraduate in

mathematics/statistics. But for the right student with the right background it's a nice way to build on senior mathematics and provide a glimpse into advanced probability. Given the right student, I would certainly try this again. If you're looking to challenge a student with theoretical material, large deviations can make a nice project.

6. Outline for "Large Deviations for I.I.D. Data with Applications to Hypothesis Testing"

6.1 Overview

Suppose $\{X_1, X_2, ...\}$ is an i.i.d. sequence with mean μ and $S_n = \sum^n_{i=1} X_n$. Then $S_n/n \rightarrow \mu$ as $n \rightarrow \infty$ by the Strong Law of Large Numbers. So, if $a > \mu$, $P(S_n \geq na) = P(S_n/n \geq a)$ must go to zero as $n \rightarrow \infty$. In many cases, $P(S_n \geq na)$ is distributed as e^{-Kn} where K is a rate that can be found. Such cases are interesting in hypothesis testing in statistics, for example. Numerous other applications exist where independence is weakened (e.g. to Markov chains).

References-Books

Patrick Billingsley, *Probability and Measure*, 2nd ed., 1986. (pp. 142-149)

James A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*, 1990. (pp. 1-14, 91-93)

Amir Dembo, Ofer Zeitouni, *Large Deviations Techniques and Applications*, 1998. (Sections 2.2.1, 3.4) (more advanced reading)

Frank den Hollander, *Large Deviations*, 2000. (Chapters 1,6- a total of 15 pages)

References-Papers

H. Chernoff. "A measure of the asymptotic efficiency of tests of a hypothesis based on the sum of observations." *Annals of Mathematical Statistics*, 23:493--507, 1952.

Alan Weiss, "An Introduction to Large Deviations for Communication Networks",*IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, August 1995 pp. 938--952. (Sections 1-5)

6.4 Weeks 1-8 - Exercises

6.4.1 Introduction

Read Bucklew pp. 1-4.

6.4.2 Week 1 - Coin Tossing

Read Weiss Section II.

1. Suppose $\{X_1,X_2,...\}$ is an i.i.d. sequence with $P(X_i = 0) = P(X_i = 1) = 0.5$ so the $E(X_i) = 0.5$. Let

$S_n = \sum^n_{i=1} X_i$. Use Stirling's approximation to $n!$ and the Binomial Theorem to show for $0.5 < x \leq 1$

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n \geq nx) = -I(x)$$

where

$$I(x) = \begin{cases} \log 2 + x \log x + (1-x)\log(1-x) & 0 \leq x \leq 1 \\ \infty & \text{otherwise.} \end{cases}$$

6.4.3 Weeks 2-3 - Chernoff's Theorem (Cramér's Theorem)

Read Billingsley pp. 142-148, Weiss III-IVc.

2. Compute the rate function (i.e. the Fenchel-Legendre transform) for these distributions. a) Bernoulli(p), b) exponential(λ), c) Poisson(10) and d) N(μ, σ^2).

3. Use a computer package to plot the rate functions for these distributions. a) Bernoulli(0.5), b) exponential(10), c) Poisson(10) and d) N(5,1)

4. Billingsley (Chapter 1, 9.2)

Let Y be a random variable with moment generating function $M(t)$. Show that $P(Y > 0) = 0$ implies that $P(Y \geq 0) = \inf_t M(t)$. Is the infimum achieved in this case?

6.4.4 Weeks 4-5 - Large Deviation Principle

Read Weiss IVd-V, Dembo & Zeitouni p.26 & Section 2.2.1

5. a) State the Large Deviations Principle with rate function $I(\cdot)$ in the Cramér setting, in terms of closed sets F and open sets G.

   b) State the Large Deviations Principle with rate function $I(\cdot)$ in the Cramér setting for a set $[a,\infty)$. (Hint: use Corollary 2.2.19).

6. The Fenchel-Legendre transform is $I(x) = \sup_\theta[\theta x - \log M(\theta)]$, where $M(\theta)$ is the moment generating function.

   a) Show $I(x)$ is non-negative.
   b) A probabilistic interpretation of Fatou's lemma states that if $\{X_1, X_2, ...\}$ is a sequence of identically distributed non-negative random variables with finite mean

   $$E[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} E[X_n].$$

   Use Fatou's lemma to show $I(x)$ is lower semi-continuous.

6.4.5 Week 6 - Application to Hypothesis Testing I

Read Billingsley pp. 148-149.

7. Verify equation 9.24 in Billingsley.
8. Consider the test following equation 9.27 in Billingsley. Imagine trying to distinguish between $p_1 = 0.2$ and $p_2 = 0.2 + t$ and also between $p_1 = 0.5$ and $p_2 = 0.5 + t$. Approximately, by what fraction can the sample size be smaller in the first case to achieve the same precision as in the second case.

6.4.6 Weeks 7-8 - Application to Hypothesis Testing II

9. Read den Hollander (Chapter 6) and complete den Hollander exercise VI.6.

OR

9. Read Chernoff (1952) and express in your own words the conclusion of Theorem 3.

6.5 Weeks 9-12 - Computer Simulation (e.g. R, S-plus or Matlab)

For this part you'll work with the N(5,1) distribution.

1. Suppose $\{X_1, X_2, ...,\}$ are i.i.d. random variables with a N(5,1) distribution, and $X_n = \sum^n_{i=1} X_i$. For any $a > 5$ what does Chernoff's theorem say about $P(S_n \geq na)$.
2. Using random number simulation, demonstrate the exponential decay and asymptotic rate described in Chernoff's theorem. That is, use simulations to approximate $P(S_n \geq na)$ and then plot it versus n for values of $a > 5$ using appropriate axes. That is, you should create a different graph for each value of a.

6.6 Weeks 13-14 - Report

Summarize the results of the exercises and computer simulation in a typed report of 8-10 pages.

References

Billingsley, P. (1986), *Probability and Measure*, Wiley Series in Probability and Mathematical Statistics, 2nd ed., John Wiley & Sons Inc., New York.

Bucklew, J. A. (1990), *Large Deviation Techniques in Decision, Simulation, and Estimation*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York.

Chernoff, H. (1952), "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, Ann. Math. Statistics, 23, 493-507.

Dembo, A. & Zeitouni, O. (1998), *Large Deviations Techniques and Applications*, Vol. 38 of *Applications of Mathematics (New York)*, Springer-Verlag, New York.

den Hollander, F. (2000), *Large Deviations*, Vol. 14 of *Fields Institute Monographs*, American Mathematical Society, Providence, RI.

Weiss, A. (1995), "An introduction to large deviations for communication networks",*IEEE Journal on Selected Areas in Communications*, 13, 938-952.

David A. Rolls
Department of Mathematics and Statistics
The University of Melbourne
Parkville, VIC 3010
Australia
D.Rolls@ms.unimelb.edu.au