I232 Information Theory
Chapter 2: Tour of Probability Theory

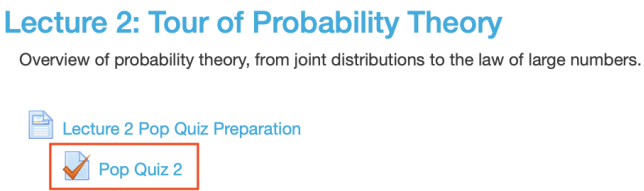
Brian Kurkoski

Japan Advanced Institute of Science and Technology

2023 April

Lecture 2 Pop Quiz

You should now take Pop Quiz 2 on the LMS.



Just for Fun

Mr. Smith has two children who are playing in a garden. We can see that one of the children is a girl, but we cannot see the gender of the other child. What is the probability both children are girls?

Assume $\Pr(G) = \Pr(B) = 0.5$.

AI-generated art in response to the prompt "Two children playing in a garden. One of the children's face is turned away from the camera. black and white, line art." starryai.com

Outline

2.1 Random Variables

Before class you studied:

2.1.1 Single Random Variables

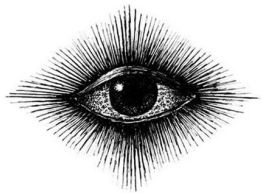
2.1.2 Jointly Distributed Random Variables

2.1.3 Conditional Probability Distributions

2.1.4 Bayes' Rule

Let's review interesting points from Section 2.1.

The "All-Knowing" Joint Distribution



Like the all-knowing eye, the **joint distribution** gives complete information about two random variables:

$$p_{\mathsf{XY}}(x,y) \longrightarrow p_{\mathsf{X}}(x), p_{\mathsf{Y}}(y) \qquad \text{by marginalization}$$
$$p_{\mathsf{XY}}(x,y) \longrightarrow p_{\mathsf{X|Y}}(x|y), p_{\mathsf{Y|X}}(y|x) \qquad \text{by definition of conditional probability}$$

Discrete Memoryless Channel

A *discrete memoryless channel* (DMC) consists of:

- an input alphabet $\mathcal{X}$,
- an output alphabet $\mathcal{Y}$ and
- a conditional probability distribution $p_{\mathsf{Y}|\mathsf{X}}(y|x)$.

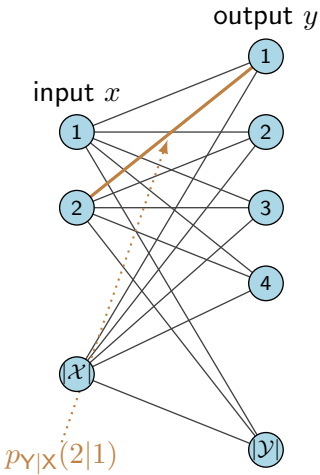
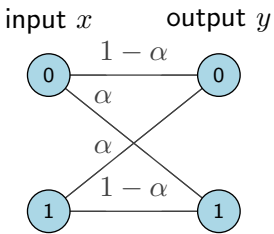


Figure 1: Transition diagram for DMC.

Binary Symmetric Channel (BSC)

In the binary symmetric channel (BSC), an error occurs with probability α.

It has binary inputs and binary outputs.



The probability transition matrix $p_{\mathsf{Y}|\mathsf{X}}(y|x)$ is:

$$p_{\mathsf{Y}|\mathsf{X}}(y|x) = \begin{bmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{bmatrix},$$

where $0 \leq \alpha \leq 1$. ★1

2.2 Independence, Expected Value, Variance

2.2.1 Independence

Definition

Two random variables X and Y are *independent* if and only if:

$$p_{\mathsf{XY}}(x, y) = p_{\mathsf{X}}(x) p_{\mathsf{Y}}(y).$$

Also, X and Y are independent if an only if

$$p_{\mathsf{X|Y}}(x|y) = p_{\mathsf{X}}(x),$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}$. This is easily obtained using $p_{\mathsf{XY}}(x, y) = p_{\mathsf{X|Y}}(x|y) p_{\mathsf{Y}}(y)$.

If $p_{\mathsf{X}}(x) = p_{\mathsf{Y}}(x)$ for all $x \in \mathcal{X}$, then we say X and Y are *independent and identically distributed*, often abbreviated iid.

2.2.2 Expected Value

Definition

The *expected value* $E[\mathsf{X}]$, of a random variable X with probability distribution $p_{\mathsf{X}}(x)$ is:

$$E[\mathsf{X}] = \sum_{x \in \mathcal{X}} x p_{\mathsf{X}}(x)$$

The expected value $E[\mathsf{X}]$ is sometimes called the mean.

The expectation of a deterministic function g of a random variable X is:

$$E[g(\mathsf{X})] = \sum_{x \in \mathcal{X}} g(x) p_{\mathsf{X}}(x)$$

★2

Variance

Definition

The *variance* of a random variable X, denoted $\mathrm{Var}[\mathsf{X}]$, is:

$$\mathrm{Var}[\mathsf{X}] = E[\mathsf{X}^2] - (E[\mathsf{X}])^2.$$

For **any** $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n$ and constants $a_1, a_2, \ldots, a_n$,

$$E[a_1\mathsf{X}_1 + a_2\mathsf{X}_2 + \cdots a_n\mathsf{X}_n] = a_1E[\mathsf{X}_1] + a_2E[\mathsf{X}_2] + \cdots a_nE[\mathsf{X}_n].$$

For **independent** $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n$:

$$\mathrm{Var}[a_1\mathsf{X}_1 + a_2\mathsf{X}_2 + \cdots a_n\mathsf{X}_n] = a_1^2\mathrm{Var}[\mathsf{X}_1] + a_2^2\mathrm{Var}[\mathsf{X}_2] + \cdots a_n^2\mathrm{Var}[\mathsf{X}_n]$$

★3

2.2.3 Events and Their Union Bound

An event is a binary random variable, which is either true or false.

An event is an outcome from an experiment, to which a probability may be assigned.

If E is an event, then $\Pr(\mathsf{E})$ is "the probability that the event E occurs."

Let $\mathsf{E}_1, \mathsf{E}_2, \ldots, \mathsf{E}_n$ be events. We want to know the probability that **at least one** occurs.

The union bound is:

$$\Pr(\bigcup_{i=1}^{n} \mathsf{E}_i) \leq \sum_{i=1}^{n} \Pr(\mathsf{E}_i)$$

Example: Union Bound of Events

Let X be the outcome of rolling a die:

$$p_{\mathsf{X}}(x) = \frac{1}{6}, x = 1, 2, 3, 4, 5, 6$$

Let:

$$\mathsf{E}_1 = \text{event } \mathsf{X} \text{ is a multiple of } 3$$
$$\mathsf{E}_2 = \text{event } \mathsf{X} \text{ is a prime number}$$

Find the union bound on $\Pr[\mathsf{E}_1 \cup \mathsf{E}_2]$ and then find it exactly.

Probability of Exclusive Events

Let E_1 and E_2 be two exclusive events. Then:

$$\Pr(\mathsf{E}_1 \cup \mathsf{E}_2) = \Pr(\mathsf{E}_1) + \Pr(\mathsf{E}_2).$$

or more generally if all $\mathsf{E}_1, \mathsf{E}_2, \ldots, \mathsf{E}_n$ are exclusive:

$$\Pr(\bigcup_{i=1}^{n} \mathsf{E}_i) = \sum_{i=1}^{n} \Pr(\mathsf{E}_i)$$

★4

2.3 Random Vectors

A random variable X_i has distribution $p_{\mathsf{X}}(x)$ on sample space $\mathcal{X}$. Then, a *random vector* $\mathbf{X}$ is a sequence of n random variables:

$$\mathbf{X} = \big(\mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3, \ldots, \mathsf{X}_n\big).$$

The sample space of $\mathbf{X}$ is the Cartesian product $\mathcal{X}^n$. ★5

Binary Random Vectors

A binary random vector $\mathbf{X}$ with parameters p and n:

$$\mathbf{X} = (\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n)$$

is a vector of n binary random variables X_i with:

$$p_{\mathsf{X}}(x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

The sample space is $\mathcal{X}^n = \{0, 1\}^n$.

Binomial Random Variable

The binary random vector $\mathbf{X}$ has z ones and $n - z$ zeros. Let Z be the sum of X:

$$\mathsf{Z} = \sum_{i=1}^{n} \mathsf{X}_i$$

so that Z is a random variable expressing the number of ones in $\mathbf{X}$.

Z is the *binomial random variable*.

Definition

A binomial random variable Z with parameters n and p has sample space $\mathcal{Z} = \{0, 1, \ldots, n\}$ and probability distribution:

$$p_{\mathsf{Z}}(z) = \binom{n}{z} p^z (1-p)^{n-z} \quad \text{for } z = 0, 1, \ldots, n,$$

Here $\binom{n}{z}$ is the number of distinct ways to place z ones into n positions.

Example

Consider an i.i.d. binary random vector with $n = 4$ and $p = \frac{1}{4}$. The sample space is:

$$\mathcal{X}^4 = \{0000, 0001, 0010, \ldots, 1111\}.$$

(a) What is the probability of $\mathbf{X} = 0110$?

(b) What is the probability that $\mathbf{X}$ has two 1's and two 0's.

(c) What is the probability that $\mathbf{X}$ has one or two 1's?

★6

2.4 Law of Large Numbers

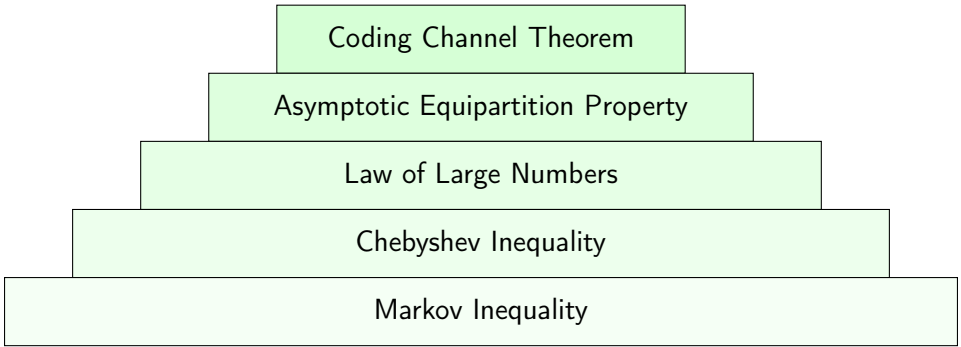
2.4.1 Markov inequality

2.4.2 Chebyshev inequality

2.4.3 Law of Large Numbers

Law of Large Numbers

"Information theory is the clever application of the law of large numbers."



The Markov inequality is used to prove the Chebyshev inequality is used to prove the law of large numbers, etc.

Sample Mean and True Mean

Definition

Let $\mathbf{X} = \mathsf{X}_1\mathsf{X}_2 \cdots, \mathsf{X}_n$ be a vector of n random variables, independent and identically distributed. The *sample mean* $\overline{\mathsf{X}}_n$ is:

$$\overline{\mathsf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathsf{X}_i$$

The sample mean $\overline{\mathsf{X}}_n$ is also a random variable.

We call $E[\mathsf{X}_i]$ the *true mean*. The sample mean $\overline{\mathsf{X}}_n$ may be close to the true mean $E[\mathsf{X}_i]$, but is not necessarily the same.

Example 1: Sample Mean vs. True Mean

Let X be uniformly distributed on $\{0, 1, 2\}$. The true mean is $E[\mathrm{X}] = 1$.

In-Class Experiment Let's generate random numbers on 0, 1, 2 by taking your birthday modulo 3. Example: December 10 $\rightarrow 10 \bmod 3 = 1$.

★Poll - Birthday Modulo 3

Example 2: Sample Mean of Binary Random Vector

Consider a binary random vector with $n = 15$ and $p = \frac{1}{3}$:

$$\mathbf{X} = \mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3, \mathsf{X}_4, \mathsf{X}_5, \mathsf{X}_6, \mathsf{X}_7, \mathsf{X}_8, \mathsf{X}_9, \mathsf{X}_{10}, \mathsf{X}_{11}, \mathsf{X}_{12}, \mathsf{X}_{13}, \mathsf{X}_{14}, \mathsf{X}_{15}$$

What is the true mean? What is the sample mean?

The true mean is clearly $\frac{1}{3}$.

Run the Experiment Many Times

1 1 1 0 1 1 0 1 0 0 0 0 1 0 0 sample mean = 0.466667

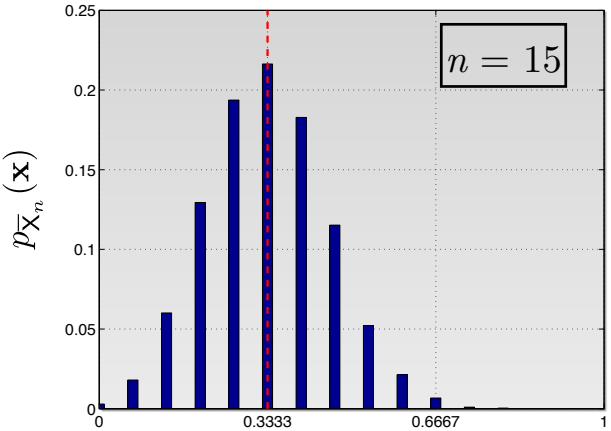
Run the Experiment Many Times

1	1	1	0	1	1	0	1	0	0	0	0	1	0	0	sample mean = 0.466667
0	0	0	1	0	0	0	1	0	1	1	0	1	1	0	sample mean = 0.4

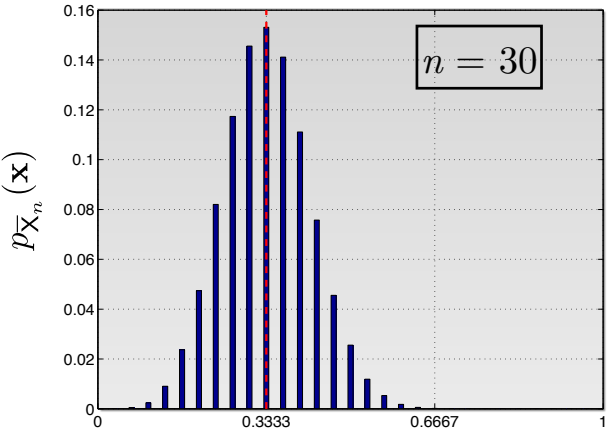
Run the Experiment Many Times

1	1	1	0	1	1	0	1	0	0	0	0	1	0	0	sample mean = 0.466667
0	0	0	1	0	0	0	1	0	1	1	0	1	1	0	sample mean = 0.4
1	1	0	0	1	1	1	0	1	1	1	1	0	0	0	sample mean = 0.6
0	0	0	1	1	0	0	0	0	0	0	1	0	1	0	sample mean = 0.266667
0	0	0	0	1	0	1	1	1	0	0	0	1	1	0	sample mean = 0.4
0	1	1	0	1	1	0	1	0	0	1	1	0	1	1	sample mean = 0.6
1	0	0	0	1	0	0	1	1	0	0	0	1	0	0	sample mean = 0.333333
1	0	1	1	0	0	1	1	1	1	0	1	0	0	0	sample mean = 0.533333
1	1	1	1	0	1	0	1	1	1	0	0	1	1	1	sample mean = 0.733333
1	1	1	1	0	1	0	0	1	0	0	1	0	0	0	sample mean = 0.466667

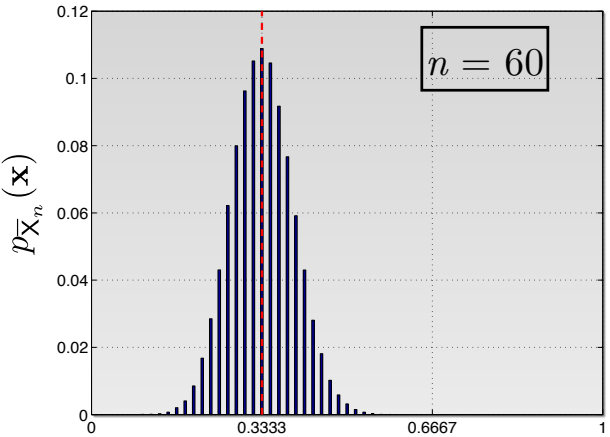
Distribution of Sample Mean When $n = 15$



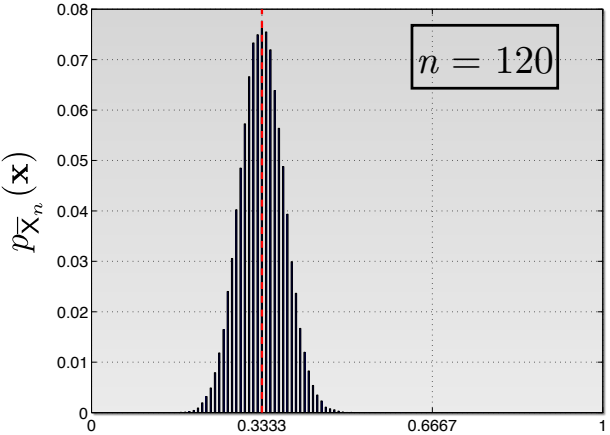
Distribution of Sample Mean When $n = 30$



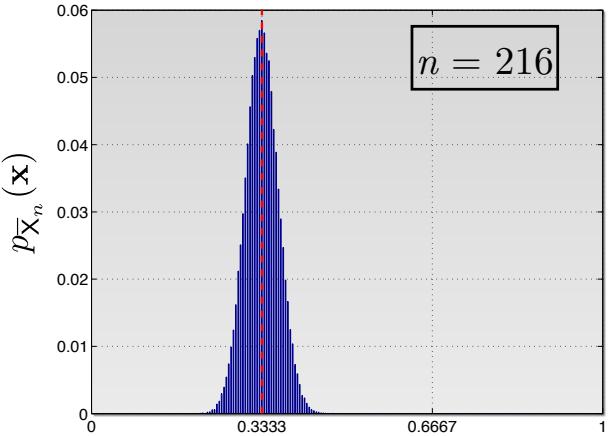
Distribution of Sample Mean When $n = 60$



Distribution of Sample Mean When $n = 120$



Distribution of Sample Mean When $n = 216$



How Close is Sample Mean to True Mean?

Sequence **X** with sample mean $\overline{\mathsf{X}}_n$ is "epsilon close" to its mean $E[\mathsf{X}]$ is expressed as:

$$|\overline{\mathsf{X}}_n - E[\mathsf{X}]| \leq \epsilon$$

Since $\overline{\mathsf{X}}_n$ is a random variable, we can consider:

$$\Pr[|\overline{\mathsf{X}}_n - E[\mathsf{X}]| \leq \epsilon]$$

The **Chebyshev Inequality** is a bound q on this probability:

$$\Pr[|\overline{\mathsf{X}}_n - E[\mathsf{X}]| \leq \epsilon] \geq q$$

★Poll - Sample Mean as n to Infinity

2.4.1 Markov inequality

Proposition

Markov Inequality If X is any nonnegative random variable and $a > 0$, then:

$$\Pr(\mathsf{X} \geq a) \leq \frac{E[\mathsf{X}]}{a}$$

2.4.2 Chebyshev inequality

What is the probability that the realization of a random variable X is close to $E[\mathsf{X}]$?

The Chebyshev inequality gives a bound on this.

Proposition

Chebyshev inequality Let X be a random variable with finite expected value $E[\mathsf{X}]$ and finite non-zero variance $\mathrm{Var}[\mathsf{X}]$. Then for $\epsilon > 0$:

$$\Pr(|\mathsf{X} - E[\mathsf{X}]| < \epsilon) \geq 1 - \frac{\mathrm{Var}[\mathsf{X}]}{\epsilon^2}.$$

Example 2 Continued: How Big Does n Need to Be?

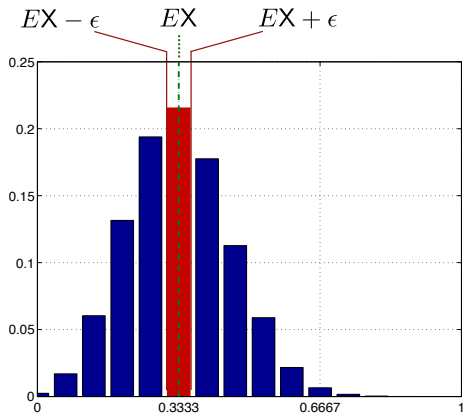
How big does n need to be such that:

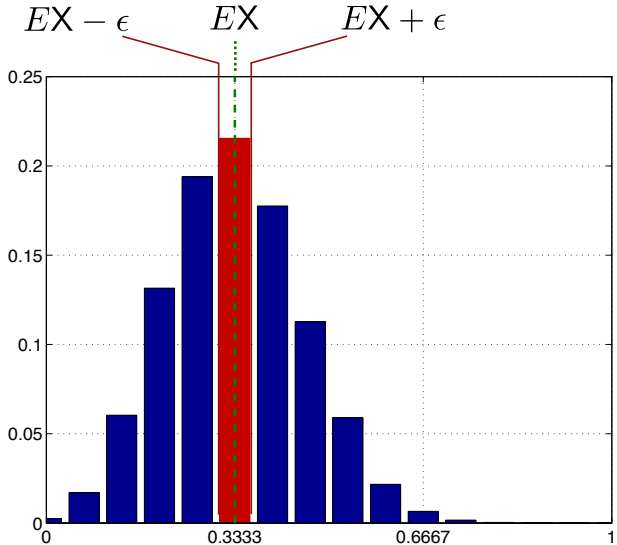
$$\Pr[|\overline{\mathsf{X}}_n - E\mathsf{X}| \leq \epsilon] > \delta$$

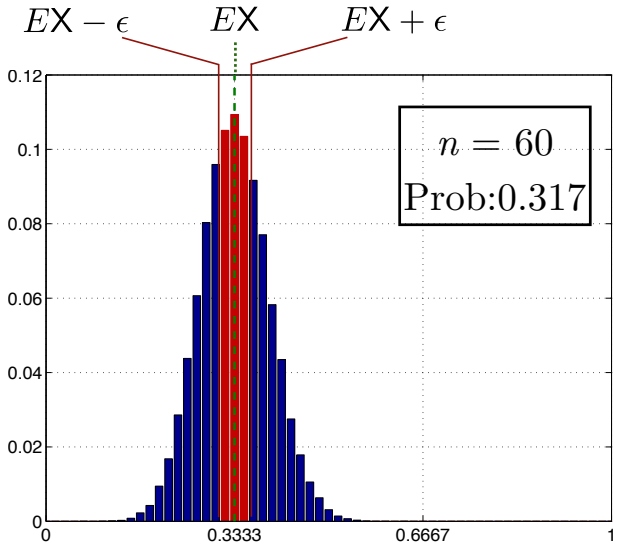
Example. What n is needed for:

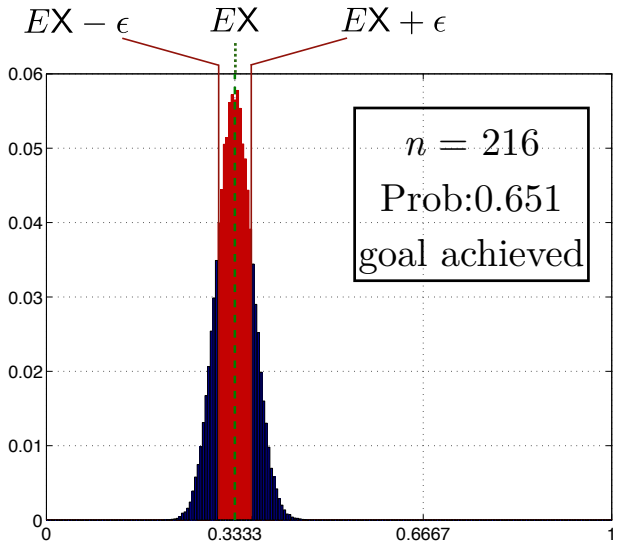
$$\epsilon = 0.03$$
$$\delta = 0.65$$

If $n = 15$, then the red area represents: $\Pr[|\overline{\mathsf{X}}_n - E\mathsf{X}| \leq \epsilon] \approx 0.25$. Must increase n to achieve $\delta = 0.65$.




$E\mathsf{X} - \epsilon$
$E\mathsf{X}$
$E\mathsf{X} + \epsilon$
0.25
0.2
0.15
0.1
0.05
0
0
0.3333
0.6667
1


$EX - \epsilon$
EX
$EX + \epsilon$
$n = 60$
Prob:0.317
0.12
0.1
0.08
0.06
0.04
0.02
0
0
0.3333
0.6667
1


$E\mathsf{X} - \epsilon$
$E\mathsf{X}$
$E\mathsf{X} + \epsilon$
0.06
0.05
0.04
0.03
0.02
0.01
0
0
0.3333
0.6667
1
$n = 216$
Prob:0.651
goal achieved

2.4.3 Law of Large Numbers

Let $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n$ be a sequence of i.i.d. random variables with mean $E[\mathsf{X}]$.

The *weak law of large numbers* states that the sample mean converges in probability towards the expected value. That is, for any positive number ε

$$\lim_{n \to \infty} \Pr\big(|\overline{\mathsf{X}}_n - E[\mathsf{X}]| < \varepsilon \big) = 1.$$

Class Info

- Tutorial hours: Monday, April 17 at 13:30. Probability exercises
- Next lecture: Wednesday, April 19 at 10:50. Mutual information and Kullback-Leibler Divergence.
- Lecture 3 Pop Quiz Preparation now available.
- Homework 1 and 2 on LMS. Deadline: Monday, April 23 at 18:00

Probability Exercise 1

Let $\mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3$ be independently and identically distributed binary random variables with sample space $\mathcal{X} = \{0, 1\}$ and $p_{\mathsf{X}}(x) = [1 - p, p]$.

(a) What is the sample space of $\mathbf{X}$, where $\mathbf{X} = \mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3$?

(b) Find the joint distribution $p_{\mathbf{X}}(\mathbf{x})$.

Probability Exercise 2

Let $\mathsf{X}_1, \mathsf{X}_2$ be independently and identically distributed random variables with sample space $\mathcal{X} = \{0, 1, 2\}$ and

$$p_{\mathsf{X}}(x) = \begin{cases} 1/2 & \text{if } x = 0 \\ 1/3 & \text{if } x = 1 \\ 1/6 & \text{if } x = 2 \end{cases} \tag{1}$$

(a) What is the sample space of $\mathbf{X}$, where $\mathbf{X} = \mathsf{X}_1, \mathsf{X}_2$,?

(b) Find the joint distribution $p_{\mathbf{X}}(\mathbf{x})$.

Probability Exercise 3

Let $\mathbf{Z}$ be a binary random vector:

$$\mathbf{Z} = (\mathsf{Z}_1, \mathsf{Z}_2, \ldots, \mathsf{Z}_n) \tag{2}$$

with $p_{\mathsf{Z}}(z) = [(1-p), p]$.

Think of Z_i as an error in position i. An error occurs with a small probability p. $\mathbf{Z}$ is an "error vector", indicating position of errors in a sequence of length n.

Assume $p = 0.1$ and $n = 8$.

(a) What is the probability of 0 errors?

(b) What is the probability of 1 error?

(c) What is the probability of 0, 1 or 2 errors?

Probability Exercise 4

Let X have sample space $\mathcal{X} = \{0, 1\}$ and be distributed as $p_{\mathsf{X}}(x) = [\frac{2}{3}, \frac{1}{3}]$. Let Z have sample space $\mathcal{Z} = \{0, 2\}$ and be distributed as $p_{\mathsf{Z}}(0) = \frac{1}{2}$ and $p_{\mathsf{Z}}(2) = \frac{1}{2}$. X and Z are independent.

Now, let $\mathsf{Y} = \mathsf{X} \cdot \mathsf{Z}$, where $\cdot$ is usual multiplication.

(a) What is the joint distribution $p_{\mathsf{XZ}}(x, z)$?

(b) What is the sample space $\mathcal{Y}$?

(c) What is the probability distribution $p_{\mathsf{Y}}(y)$?

(d) What are the probability distributions $p_{\mathsf{Y}|\mathsf{Z}}(y|0)$ and $p_{\mathsf{Y}|\mathsf{Z}}(y|2)$.

(e) What is the entropy $H(\mathsf{X})$ and $H(\mathsf{Z})$?

(f) What is the entropy $H(\mathsf{Y}|\mathsf{Z})$?