

Introduction to Data Science With Probability and Statistics

Lecture 19: The Law of Large Numbers

CSCI 3022 - Summer 2020

Sourav Chakraborty

Dept. of Computer Science

University of Colorado Boulder

The law of large numbers

- It is observed that two measurements concerning natural phenomena, results in two different outcomes even when performed in seemingly identical conditions.
 - To overcome this, we perform repeated measurements and we take the average of all the outcomes.
 - Each of those measurements are independent random variables with their own unknown distribution.
- It is a probabilistic fact that from such a sequence—in principle—any feature of the distribution can be recovered. This is a *consequence* of the law of large numbers.



Averages vary less

Let us consider a sequence of random variables $X_1, X_2, X_3, \dots, X_n$ having identical distributions.

We shall denote the distribution function of each random variable X_i by F , its expectation by μ , and the standard deviation by σ .

Average of first n random variables,
$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

EXPECTATION AND VARIANCE OF AN AVERAGE. If \bar{X}_n is the average of n independent random variables with the same expectation μ and variance σ^2 , then

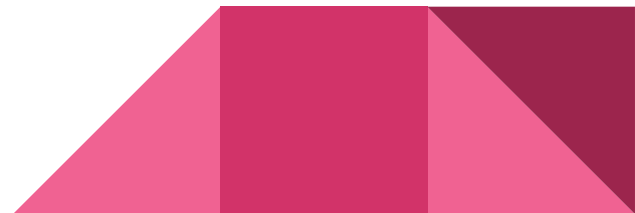
$$E[\bar{X}_n] = \mu \quad \text{and} \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Chebyshev's inequality

Chebyshev's inequality. For an arbitrary random variable Y and any $a > 0$:

$$P(|Y - E[Y]| \geq a) \leq \frac{1}{a^2} \text{Var}(Y).$$

THE “ $\mu \pm \text{A FEW } \sigma$ ” RULE. Most of the probability mass of a random variable is within a few standard deviations from its expectation.



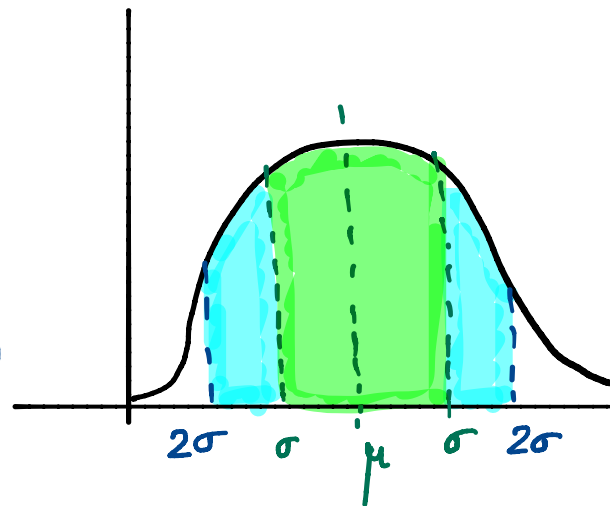
Chebyshev's inequality

Chebyshev's inequality. For an arbitrary random variable Y and any $a > 0$:

$$P(|Y - E[Y]| \geq a) \leq \frac{1}{a^2} \text{Var}(Y).$$

$$\begin{aligned} P(|Y - \mu| < k\sigma) &= 1 - P(|Y - \mu| \geq k\sigma) \\ &\geq 1 - \frac{1}{(k\sigma)^2} \text{Var}(Y) \end{aligned}$$

$$P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$



Chebyshev's inequality

Calculate $P(|Y - \mu| < k\sigma)$ exactly for $k = 1, 2, 3, 4$ when Y has an $\text{Exp}(1)$ distribution and compare this with the bounds from Chebyshev's inequality

Imp. info: $\mu = \frac{1}{\lambda}$, $\text{Var} = \frac{1}{\lambda^2}$ for $\text{Exp}(1)$

$$\text{CDF} : P(X \leq a) = 1 - e^{-a}$$

for values' calculation can use calc. or python.



Chebyshev's inequality

Calculate $P(|Y - \mu| < k\sigma)$ exactly for $k = 1, 2, 3, 4$ when Y has an $\text{Exp}(1)$ distribution and compare this with the bounds from Chebyshev's inequality

$$\begin{aligned} P(|Y - \mu| < k\sigma) &= P(|Y - 1| < k) = P(1 - k < Y < k + 1) \\ &= \underbrace{P(Y < k + 1)}_{\text{CDF}} = 1 - e^{-k-1} \end{aligned}$$

k	1	2	3	4
lower bound	0	0.750	0.889	0.938
Exact value	0.865	0.950	0.982	0.993

The law of large numbers

THE LAW OF LARGE NUMBERS. If \bar{X}_n is the average of n independent random variables with expectation μ and variance σ^2 , then for any $\varepsilon > 0$:

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0.$$

There is a stronger law of large numbers -> Even though it is a strong statement, the law of large numbers in this paragraph is more accurately known as the weak law of large numbers.

A stronger result holds, the strong law of large numbers, which says that:

$$P\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mu\right) = 1.$$

New Material

You are trying to determine the melting point of a new material, of which you have a large number of samples. For each sample that you measure you find a value close to the actual melting point c but corrupted with a measurement error. We model this with random variables:

$$M_i = c + U_i$$

where M_i is the measured value in degree Kelvin, and U_i is the occurring random error. It is known that $E[U_i] = 0$ and $Var(U_i) = 3$ for each i , and that we may consider the random variables M_1, M_2, \dots independent. According to Chebyshev's inequality, how many samples do you need to measure to be 90% sure that the average of the measurements is within half a degree of c ?



New Material

We want: $P(|\bar{M}_n - c| \leq 0.5) \geq 0.9$, where n : total number of readings

$$\text{So, } P(|U_n| \leq 0.5) \geq 0.9$$

$$P(|U_n| > 0.5) \leq \frac{1}{(0.5)^2} \frac{\text{Var}(U_i)}{n} = \frac{12}{n}$$

$$\text{So; } \frac{12}{n} \leq 0.1 \Rightarrow \boxed{n \geq 120}$$

\bar{M}_n : mean of the n readings.

Consequences of the law of large numbers


➤ Recovering the probability of an event

Suppose we want to find the probability of an event such as: $p = P(X \in C)$, where $C = (a, b]$ for some $a < b$.

$$Y_i = \begin{cases} 1 & \text{if } X_i \in C, \\ 0 & \text{if } X_i \notin C. \end{cases} \quad \text{Here, random variable } Y_i \text{ is called an } \textit{indicator random variable}$$

$$E[Y_i] = 1 \cdot P(X_i \in C) + 0 \cdot P(X_i \notin C) = P(X_i \in C) = P(X \in C) = p$$

$$E[\bar{Y}_n] = E\left[\sum Y_i / n\right] = \frac{n \cdot p}{n} = p.$$


$$\lim_{n \rightarrow \infty} P(|Y_n - p| > \varepsilon) = 0 \quad \text{for any } \varepsilon > 0.$$

Random Samples and Statistical Models

- Michelson conducted an experiment between June 5 and July 2 in 1879 where he recorded 100 measurements of speed of light as shown in the table.
- Each of these measurement is an independent random variable as they were all recorded under identical conditions
- Hence, it is justified to assume their probability distributions are also same

$$X_1, X_2, \dots, X_{100}$$

850	740	900	1070	930	850	950	980	980	880
1000	980	930	650	760	810	1000	1000	960	960
960	940	960	940	880	800	850	880	900	840
830	790	810	880	880	830	800	790	760	800
880	880	880	860	720	720	620	860	970	950
880	910	850	870	840	840	850	840	840	840
890	810	810	820	800	770	760	740	750	760
910	920	890	860	880	720	840	850	850	780
890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

in km/sec minus 299000

Random Samples and Statistical Models

RANDOM SAMPLE. A *random sample* is a collection of random variables X_1, X_2, \dots, X_n , that have the same probability distribution and are mutually independent.

STATISTICAL MODEL FOR REPEATED MEASUREMENTS. A dataset consisting of values x_1, x_2, \dots, x_n of repeated measurements of the same quantity is modeled as the realization of a random sample X_1, X_2, \dots, X_n . The model may include a partial specification of the probability distribution of each X_i .



Random Samples and Statistical Models

We obtain a dataset of ten elements by tossing a coin ten times and recording the result of each toss. What is an appropriate statistical model and corresponding model distribution for this dataset?

Imp. Info : we are modelling a real life scenario;

So,

→ "true" distribution + parameters are real life ones.

→ model distribution + parameter are (estimated ones) of the model



Random Samples and Statistical Models

We obtain a dataset of ten elements by tossing a coin ten times and recording the result of each toss. What is an appropriate statistical model and corresponding model distribution for this dataset?

If its a fair coin : It can be modelled as $\text{Ber}(1/2)$, coincides the true distribution & parameters.

Else : $\text{Ber}(p)$, where we don't know p . We will have to estimate.



Random Samples and Statistical Models

Important Questions to ask about a statistical model-

- *Which feature of the model distribution* represents the quantity of interest and *how do we use our dataset* to determine a value for this?
- *Which model distribution* fits a particular dataset best?



Distribution features and sample statistics

We know that empirical summaries of datasets can be represented as a function:

$$h(x_1, x_2, \dots, x_n)$$

Since datasets are modeled as realizations of random samples X_1, X_2, \dots, X_n , an object $h(x_1, x_2, \dots, x_n)$ is a realization of the corresponding random object:

$$h(X_1, X_2, \dots, X_n)$$

Such an object, which depends on the random sample X_1, X_2, \dots, X_n only, is called a sample statistic.

For the sample statistic, $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ based on a sample X_1, X_2, \dots, X_n from a probability distribution with expectation μ .

Law of large number states: $\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$ for every $\varepsilon > 0$.

Estimating features of the “true” distribution

Sample statistic	Distribution feature
Graphical	
Empirical distribution function F_n	Distribution function F
Kernel density estimate $f_{n,h}$ and histogram	Probability density f
(Number of X_i equal to a)/ n	Probability mass function $p(a)$
Numerical	
Sample mean \bar{X}_n	Expectation μ
Sample median $\text{Med}(X_1, X_2, \dots, X_n)$	Median $q_{0.5} = F^{\text{inv}}(0.5)$
p th empirical quantile $q_n(p)$	100 p th percentile $q_p = F^{\text{inv}}(p)$
Sample variance S_n^2	Variance σ^2
Sample standard deviation S_n	Standard deviation σ
$\text{MAD}(X_1, X_2, \dots, X_n)$	$F^{\text{inv}}(0.75) - F^{\text{inv}}(0.5)$, for symmetric F

Next:

1: Kernel Density Estimate.

2: Bootstrap.

