DISCRETE PROBABILITY THEORY

DISCRETE PROBABILITY

SAMPLE SPACES AND EVENTS

DEFINITIONS

Sample Space: Sample space Ω is the set of all possible outcomes of a random experiment.

In discrete probability theory, we assume that the space Ω is finite or countably infinite.

Power Set: Let $\mathcal{F}$ denote the set of all subsets ("power set") of Ω. Recall that $\emptyset \epsilon \mathcal{F}$ and $\Omega \epsilon \mathcal{F}$.

Event: Any element $E \epsilon \mathcal{F}$ is called an event.

EXAMPLE

Consider the experiment of tossing a coin three times. What can we say about the sample space, power set, and the events "two or more heads", "no tails", and "less than two heads"?

_ **The sample space is**
$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$.

_ **The power set $\mathcal{F}$ has $2^8 = 256$ elements!**

_ **The event "two or more heads" is $E = \{HHH, HHT, HTH, THH\}$.**

_ **The event "no tails" is $F = \{HHH\}$.**

_ **And the event "less than two heads" is the complement of event E, i.e.**
$E^C = \Omega \setminus E = \{HTT, THT, TTH, TTT\}$.

AXIOMS OF DISCRETE PROBABILITY

DEFINITIONS

Discrete Probability Measure: Let Ω be a finite or countably infinite sample space with power set $\mathcal{F}$. A function $P: \mathcal{F} \to \mathbb{R}$ is called a discrete probability measure if it satisfies the following conditions:
(1) $0 \leq P(E) \leq 1$ for any event $E \epsilon \mathcal{F}$.

(2) $P(\Omega) = 1$.

(3) Finite additivity: For any collection of pairwise disjoint events, $\{E_i\}_{i \epsilon I} \subset \mathcal{F}$, $P(\bigcup_{i \epsilon I} E_i) = \sum_{i \epsilon I} P(E_i)$.

Discrete Probability Space: We say that the event $E \epsilon \mathcal{F}$ occurs with probability $P(E)$. The triple $(\Omega, \mathcal{F}, P)$ is called a discrete probability space.

CONSEQUENCES

The following properties are an immediate consequence of these axioms.

Derived Probability Rules: Let $(\Omega, \mathcal{F}, P)$ be a probability space. If $E, F \epsilon \mathcal{F}$, then

(1) $P(E^c) = 1 - P(E)$;

(2) If $E \subset F$, then $P(E) \leq P(F)$;

(3) $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

THREE DEFINITIONS OF PROBABILITY

Objective Probability: There are n equally likely outcomes of a random process and the event A consists of exactly m of these outcomes, then the probability of the event A is: $P(A) = \frac{m}{n}$.

Empirical Probability: This probability is derived through an experiment. If m_E is the number of times the event A occurs in a repeated trial and n_E the total number of trials in a random experiment, then probability of the event A is: $P(A) = \frac{m_E}{n_E}$.

Note, that if the number of trials in the experiment is large enough, the empirical probability should approach the objective probability.

Subjective Probability: An individual opinion or belief about the probability of occurrence of an event.

CONJUNCTION FALLACY

A famous error in human decision-making occurs through a violation of derived probability rule (2).

EXAMPLE: WHAT IS YOUR SUBJECTIVE PROBABILITY?

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

(Source: Kahneman, Daniel, and Amos Tversky, 1983, Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment, Psychological Review 4, 293-315.)

QUESTION

Which of the following eight options is more likely for Linda to be?

1. **Linda is a teacher in elementary school.**
2. **Linda works in a bookstore and takes yoga classes.**
3. **Linda is active in the feminist movement.**
4. **Linda is a psychiatric social worker.**
5. **Linda is a member of the 'League of Women Voters'.**
6. **Linda is a bank teller.**
7. **Linda is an insurance salesperson.**
8. **Linda is a bank teller and active in a feminist movement.**

BENFORD'S LAW

(Source: http://mathworld.wolfram.com/BenfordsLaw.html)

DEFINITION

Benford's Law: Benford's Law - also called first digit law, first digit phenomenon, or leading digit phenomenon - states that in listings, tables of statistics, etc., the digit 1 tends to occur with probability ~30%, much greater than the expected 11.1% (= 1/9 , i.e., one digit out of 9).

TABLE OF FREQUENCIES

Frequency of d being the first digit is $P(d) = log_{10}(1 + \frac{1}{d}), d \in \{1, 2, 3, \ldots, 9\}$

d	1	2	3	4	5	6	7	8	9
P(d)	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

EXAMPLES

Benford's Law has been found to hold for:

_ **credit card bills**
_ **stock market prices**
_ **market values of listed companies**
_ **populations of cities or districts**

BENFORD'S LAW IN REALITY

Benford's Law can be observed, for instance, by examining tables of logarithms and noting that the first pages are much more worn and smudged than later pages (Newcomb 1881).

Furthermore, Benford's law is often used as an indicator of fraudulent data: In 1993 Wayne James Nelson was accused and found guilty of trying to defraud the state of Arizona of nearly $2 million. Given a dataset of 23 checks issued, they found inconsistency in the data. One major flaw of of Wayne James was to not assume that some digits occur more frequently and hence the numbers didn't follow Benford's law. For more information on this case visit http://www.journalofaccountancy.com/issues/1999/may/nigrini.html.

INDEPENDENCE

DEFINITIONS

Let $(\Omega,\ \mathcal{F},\ P)$ be a probability space.

Independence for two events: Two events $E,\ F\ \epsilon\ \mathcal{F}$ are called independent if $P\left(E\cap F\right)=P(E)\times P(F)$.

Independence for a collection of events:
A collection of events, $\{E_i\}_{i\epsilon\ I}$, is said to be independent, if for any $m\ \geq\ 2$ and any choice of distinct indices $i_1,i_2,\ldots,i_m$,

$$P\left(\bigcap_{k=1}^{m}E_{i_k}\right)=\prod_{k=1}^{m}P\left(E_{i_k}\right).$$

EXAMPLE

Recall the experiment of tossing a coin three times. The sample space is $\Omega\ =\{HHH,\ HHT,\ HTH,\ THH,\ HTT,\ THT,\ TTH,\ TTT\}$.

Let us assume that the coin is "fair", so $P\left(H\right)\ =\ P(T)\ =\ \frac{1}{2}$.

- Then $P(HHH)=\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{8}\ =\ 0.125\ =\ 12.5\ \%$.
- Similarly, $P\left(HHT\right)=P(HTH)=\ \cdots=P(TTT)=0.125$.
- And for the event "exactly two heads", $G\ =\{HHT,\ HTH,\ THH\}$: $P(G)=P(HHT)+P(HTH)+P(THH)=0.375$.

CONDITIONAL PROBABILITY

UPDATING PROBABILITIES

MOTIVATION

Consider an event E with probability $P(E)$. How does the probability of E change when we know that some other event $F \subset \Omega$ has occurred? Given that the outcome lies in the set F, the event E will occur if and only if $E \cap F$ occurs. The relative chance of E occurring after receiving the information that F has occurred is, therefore, $P(E \cap F) / P(F)$. This motivates the following definition.

DEFINITION

Conditional Probability: Let $(\Omega, \mathcal{F}, P)$ be a probability space. Consider two events $E, F \in \mathcal{F}$ with $P(F) > 0$. The probability of E given that F has occurred is denoted $P(E \mid F)$ and is defined as $P(E \mid F) = \frac{P(E \cap F)}{P(F)}$.

We call $P(E \mid F)$ the conditional probability of E given F.

CONSEQUENCES

If two events $E, F \in \mathcal{F}$ with $P(E), P(F) > 0$ are independent, then $P(E \mid F) = P(E)$ and $P(F \mid E) = P(F)$.

From the definitions: $P(E \mid F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E)\,P(F)}{P(F)} = P(E)$.

TREES AND TABLES: SUCCESS OF TV ADVERTISING

INTRODUCTION

(Source: Foster, Dean and Robert Stine, 2011, Statistics for Business: Decision Making and Analysis, 1. Edition, Boston: Addison-Wesley.)

Assume there are three programs that can be viewed on a Sunday evening. Viewers can either watch '60 Minutes', 'Desperate Housewives' or a football match. We want to investigate how successful TV advertisement is given the TV program that can be watched. For each program, we collect data on the percentage of viewers that:

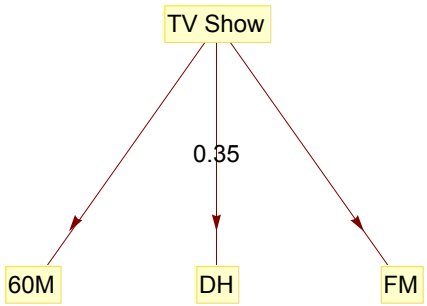
- **Watch the ads,**
- **Skip the ads.**

MARGINAL PROBABILITIES

It is given that the viewers watch the three respective shows with the following probabilities:

- $P(60\ Minutes) = 0.15,$
- $P(Desperate\ Housewives) = 0.35,$
- $P(Football\ Match) = 0.5.$

The three probabilities above represent marginal probabilities. These probabilities are illustrated in the probability tree below, where 60M stands for '60 Minutes', DH for 'Desperate Housewives' and FM for a football match resp.

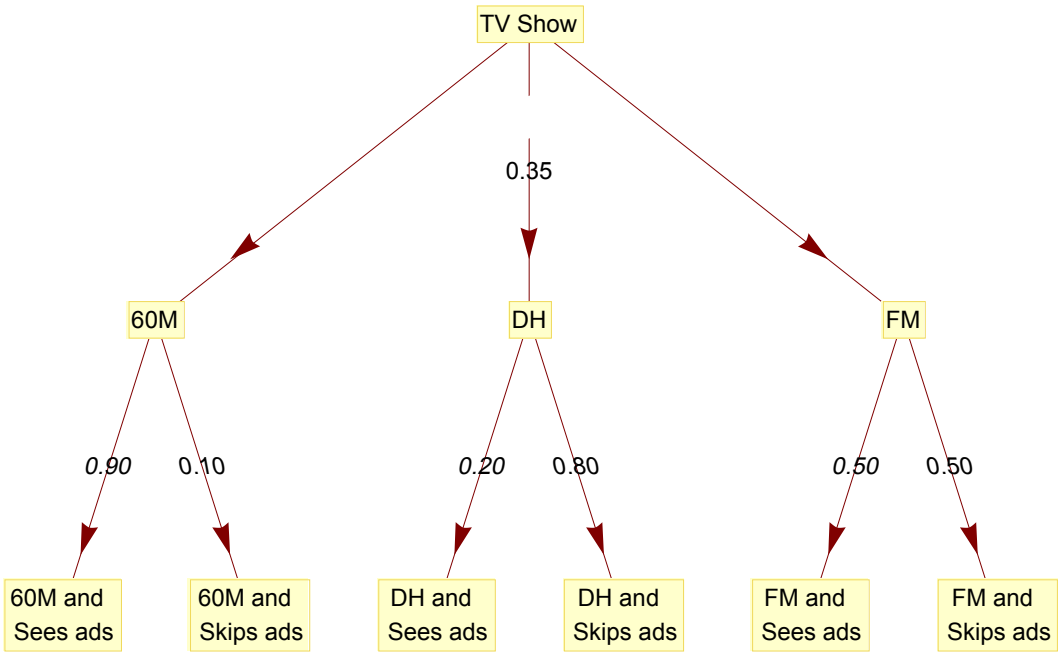


CONDITIONAL PROBABILITIES

There are six different conditional probabilities on whether a viewer watches ads or not, e.g.:

- $P(Sees\ ads \mid Football\ Match) = 0.5,$
- $P(Skips\ ads \mid Football\ Match) = 0.5,$
- ...
- $P(Skips\ ads \mid 60\ M) = 0.1.$

These conditional probabilities are illustrated in the probability tree below.



QUESTION

Using the probability tree above, what is the joint probability for 'Football match' and 'Sees Ads' (e.g. $P(Sees\ Ads \cap Football\ match)$) and what is the marginal probability for 'Sees Ads' (e.g. $P(Sees\ Ads)$)?

Solution

- $P(Sees\ Ads \cap Football\ match)$.

  We know from the definition of conditional probability that

  $P(Sees\ Ads \cap Football\ match) = P(Sees\ Ads \mid Football\ match) \cdot P(Football\ match)$.

  It is given that $P(Sees\ Ads \mid Football\ match) = 0.5$ **and** $P(Football\ match) = 0.5$. **Therefore**

  $P(Sees\ Ads \cap Football\ match) = 0.5 \cdot 0.5 = 0.25$.

- $P(Sees\ Ads)$.

  We know from 'working with probability trees' that we can calculate $P(Sees\ Ads)$ **the following way:**

  $P(Sees\ Ads) = P(60\ Min \cap Sees\ Ads) + P(Desperate\ Housewives \cap Sees\ Ads) + P(Football\ match \cap Sees\ Ads)$

We get:

$$P(Sees\ Ads) = 0.15 \cdot 0.9 + 0.35 \cdot 0.2 + 0.5 \cdot 0.5 = 0.455.$$

QUESTION

Consider the previous question. Given that a viewer is watching ads, which TV program is (s)he watching? Fill in the probability table below.

	60 Minutes	Desperate Housewives	Football match	Total
Sees Ads				
Skips Ads				
Total				1.0

Solution

From the previous question, we know that $P(Sees\ Ads) = 0.455$ and $P(Sees\ Ads \cap Football\ match) = 0.25$. Therefore we can calculate $P(Skips\ Ads)$ and $P(Skips\ Ads \cap Football\ match)$. We get:

- $P(Skips\ Ads) = 1 - 0.455 = 0.545$,
- $P(Skips\ Ads \cap Football\ match) = P(Football\ match) - P(Football\ match \cap Sees\ Ads) = 0.5 - 0.25 = 0.25$.

From the probability tree, we know that

- $P(60\ Minutes) = 0.15$,
- $P(Desperate\ Housewives) = 0.35$ **and**
- $P(Football\ match) = 0.50$.

Similar reasoning as in the previous question for the other joint probabilities yields the probability table below.

	60 Minutes	Desperate Housewives	Football match	Total
Sees Ads	0.135	0.07	0.25	0.455
Skips Ads	0.015	0.28	0.25	0.545
Total	0.15	0.35	0.50	1.0

Now we can answer the question: Given that a viewer is watching ads, which TV program is (s)he watching?

As can be seen easily in the probability table above, people who watch football are more likely to watch the ads:

$$P(Football\ match \mid Sees\ Ads) = \frac{0.25}{0.455} \approx 0.549$$

which is bigger than

$$P(60\ Minutes \mid Sees\ Ads) = \frac{0.135}{0.455} \approx 0.297$$

or

$$P(Desperate\ Housewives \mid Sees\ Ads) = \frac{0.07}{0.455} \approx 0.154.$$

BAYES' RULE

MOTIVATION

Observe that in the previous example we used knowledge about $P(F \mid E)$ to compute $P(E \mid F)$. This calculation was an application of a famous rule in probability theory, Bayes' Rule.

DEFINITION

Bayes' Rule: Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\{E_i\}_{i=1}^{n} \subset \mathcal{F}$ be a collection of mutually exclusive and collectively exhaustive subsets of Ω. If $P(F) > 0$ and $P(E_i) > 0$ for each $i \epsilon \{1, \ldots, n\}$, then

$$P(E_i \mid F) = \frac{P(F|E_i)\,P(E_i)}{\sum_{j=1}^{n} P(F|E_j)\,P(E_j)}.$$

APPLICATION: SPAM FILTER

INTRODUCTION

(Source: Foster, Dean and Robert Stine, 2011, Statistics for Business: Decision Making and Analysis, 1. Edition, Boston: Addison-Wesley.)

Assume workers of a company want to filter out junk mail from important mail messages. They base their method on past data. For example, $20\ \%$ of all emails that were considered as junk mail contained the word combination 'Nigerian general'. Past data indicates the following probabilities:

- $P(Nigerian\ general\ appears \mid Junk\ mail) = 0.20,$
- $P(Nigerian\ general\ appears \mid Not\ junk\ mail) = 0.001,$
- $P(Junk\ mail) = 0.50.$

QUESTION

Fill in the probability table below.

	Junk mail	Not junk mail	Total
Nigerian general appears			
Nigerian general does not appear			
Total			1.0

Solution

- **It is given that** $P(Junk\ mail) = 0.5$**. Therefore** $P(Not\ junk\ mail) = 1 - 0.5 = 0.5.$
- $P(Nigerian\ general\ appears \cap Junk\ mail) = P(Nigerian\ general\ appears \mid Junk\ mail) \cdot P(Junk\ mail) = 0.20 \cdot 0.50 = 0.10.$
- $P(Nigerian\ general\ appears \cap Not\ junk\ mail) = P(Nigerian\ general\ appears \mid Not\ junk\ mail) \cdot P(Not\ junk\ mail) = 0.001 \cdot 0.50 = 0.0005.$
- **Similar reasoning for** $P(Nigerian\ general\ does\ not\ appear \cap Junk\ Mail)$ **and** $P(Nigerian\ general\ does\ not\ appear \cap Not\ junk\ mail)$ **leads to the probability table below.**

	Junk mail	Not junk mail	Total
Nigerian general appears	0.1	0.0005	0.1005
Nigerian general does not appear	0.4	0.4995	0.8995
Total	0.5	0.5	1.0

QUESTION

Using the probability table above, calculate the probability that an email should be considered junk mail given that the phrase “Nigerian general” appears.

Solution

$$P(Junk\ mail \mid Nigerian\ general\ appears) = \frac{P(Junk\ mail \cap Nigerian\ general\ appears)}{P(Nigerian\ general\ appears)} = \frac{0.1}{0.1005} = 0.995.$$

We can conclude that email messages to this employee with the phrase ‘Nigerian general’ have a high probability (more than $99\ \%$) of being spam. The spam filter should move emails containing this phrase straight to the junk folder.

PROBABILITY CONFUSION — THE SAD STORY OF SALLY CLARK

Sally Clark was found guilty of murder after two of her children died. The key reasoning underlying the verdict was based on the following two statements.

- **The probability that a child dies of Sudden Infant Death Syndrome SIDS) is equal to** $\frac{1}{8543}$**.**

- **The probability that two children in the same family die of Sudden Infant Death Syndrome is equal to $\frac{1}{8543} \cdot \frac{1}{8543} \approx \frac{1}{73\ million}$, assuming that the death of one child is independent of the death of the other child within one family.**

Therefore Sally Clark was claimed to be innocent with a probability of $\frac{1}{73\ million}$.

The reasoning contained (at least) three fundamental flaws.

DISCRETE RANDOM VARIABLES

BASICS

DEFINITION

Discrete Random Variable: Let $(\Omega, \mathcal{F}, P)$ be a discrete probability space. A discrete random variable is a function $X : \Omega \to \mathbb{R}$.

Note that for a discrete probability space the image of Ω under the function $X : \Omega \to \mathbb{R}$, the set $X(\Omega) = \{x \in \mathbb{R} \mid \exists\, \omega\ x = X(\omega)\}$, is finite or countably infinite.

EXAMPLE

Recall the experiment of tossing a coin three times with the sample space $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$.

Define the random variable X to be the number of coin tosses coming up H. The image of Ω under the function $X : \Omega \to \mathbb{R}$ is $X(\Omega) = \{0, 1, 2, 3\}$. For example, $X(HHH) = 3,\ X(THT) = 1$.

DEFINITIONS

Preimage: For any function $f : A \to B$ and for any subset $S \subset B$, we define the preimage of S to be the set $f^{-1}(S) = \{a \in A \mid f(a) \in S\}$.

For a single element $b \in B$ we often omit the set braces and write $f^{-1}(b)$ instead of $f^{-1}(\{b\})$.

Let X be a discrete random variable on the probability space $(\Omega, \mathcal{F}, P)$.
For any $a \in \mathbb{R}$, we define the event $X = a$ as the event $X^{-1}(a) = \{\omega \in \Omega \mid X(\omega) = a\} \subset \mathcal{F}$.

Probability Mass Function (PMF): The PMF of X is the function $p : \mathbb{R} \to [0, 1]$ defined by $p(a) = P(X = a) = P(X^{-1}(a))$.
Since X only takes on values in $X^{-1}(\Omega)$, it holds that $\sum_{a \in X(\Omega)} p(a) = 1$.

Cumulative Distribution Function (CDF): The CDF $F : \mathbb{R} \to [0, 1]$ of X is

defined by $F(b) = P(X \leq b) = \sum_{a \leq b} p(a)$ for any $b \in \mathbb{R}$. The value $F(b)$ denotes the probability that the random variable X takes on a value of at most b.

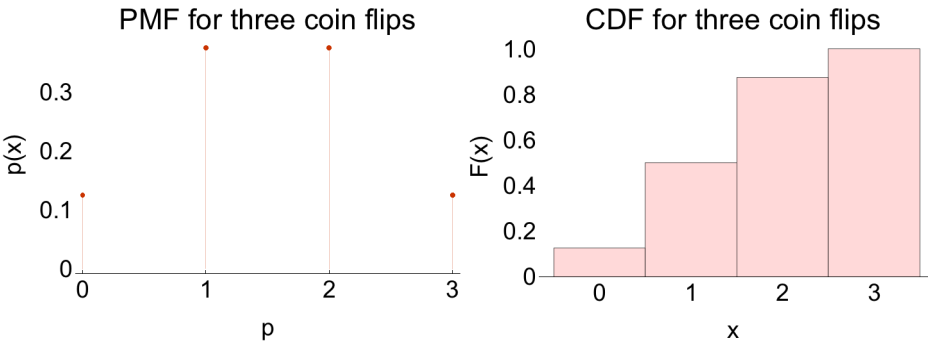
EXAMPLE

Consider again the example with the three coin flips, $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$, with the random variable X being the number of coin tosses coming up H.

Then:

- $p(0) = P(X = 0) = P\left(X^{-1}(0)\right) = P(\{TTT\}) = 0.125$;
- $p(1) =$ $P(X = 1) = P\left(X^{-1}(1)\right) = P(\{HTT, THT, TTH\}) = 0.375$;
- $p(2) =$ $P(X = 2) = P\left(X^{-1}(2)\right) = P(\{HHT, HTH, THH\}) = 0.375$;
- $p(3) = P(X = 3) = P\left(X^{-1}(3)\right) = P(\{HHH\}) = 0.125$;
- $F(-0.001) = 0$;
- $F(0) = 0.125$;
- $F(1) = 0.5$;
- $F(1.5) = 0.5$;
- $F(2) = 0.875$;
- $F(3) = 1$;
- $F(\pi) = 1$.

Graphically this can be illustrated as follows:



EXPECTATION OF A DISCRETE RANDOM VARIABLE

DEFINITION

Let X be a discrete random variable on the probability space $(\Omega, \mathcal{F}, P)$ with probability mass function p. The expected value of X, denoted $\mu = E[X]$, is given by $\mu = E[X] = \sum_{\{x:p(x)>0\}} x\, p(x) = \sum_{\{x \in X(\Omega)\}} x\, p(x)$.

We use the terms “expected value”, “mean”, and “average” synonymously. The expected value of a discrete random variable is the probability-weighted sum of all possible values.

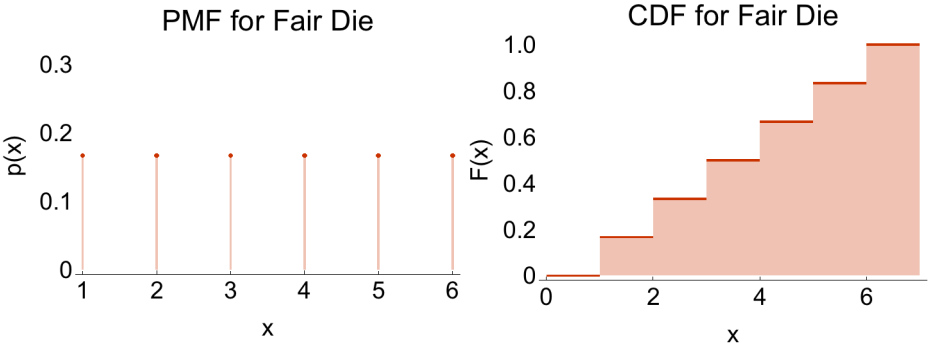
EXAMPLE

Imaging rolling a fair die once. The number of possible outcomes for throwing a fair die is obviously finite. Let X be the random variable representing the possible outcomes of throwing a fair die, then:

- **The number of possible outcomes for X is clearly finite. Therefore X is a discrete random variable.**
- **All outcomes are equally likely. Therefore X is uniformly distributed.**

  Since X is a discrete random variable, X follows the discrete uniform distribution.

In the figures below, the probability mass function and the cumulative distribution function of X are shown.



ANOTATION

The term “expected value” may sound misleading, since, as the previous example illustrates, the probability of the random variable taking on its mean may be zero. So, we may not “expect” to observe the expected value in a single random experiment.

The common interpretation of the expected value of a random variable is that it is the (statistical) average of many repetitions of the random experiment underlying the random variable. Using Monte-Carlo simulation, we can easily illustrate this interpretation for a fair die.

```
N[Mean[RandomInteger[{1, 6}, 10000]]]
```

```
3.5061
```

DEFINITION

Let X be a discrete random variable on the probability space $(\Omega, \mathcal{F}, P)$ with probability mass function p. Then for any function $g : X(\Omega) \to \mathbb{R}$ the expected value of $g(X)$, denoted $E[g(X)]$, is given by
$E[g(X)] = \sum_{\{x:p(x)>0\}} g(x)\, p(x) = \sum_{\{x \in X(\Omega)\}} g(x)\, p(x)$.

For constants $a, c \in \mathbb{R}$, $E[aX + c] = aE[X] + c$.

EXAMPLE

Recall the random variable X from a single roll of a fair die. Consider $g(X) = X^2$. Then
$E[X^2] = \sum_{i=1}^{6} i^2 P(i) = \frac{(1+4+9+16+25+36)}{6} = 91/6 \approx 15.1667$. Observe that $E[X^2] \neq (E[X])^2$.

VARIANCE OF A DISCRETE RANDOM VARIABLE

DEFINITIONS

Let X be a discrete random variable on the probability space (Ω,$\mathcal{F}$,P) with probability mass function p and expected value $\mu = E[X]$.

Variance: Then the variance of X, denoted by $\sigma^2 = Var[X]$, is given by
$Var[X] = E[(X-\mu)^2] = \sum_{\{x:\, p(x)>0\}} (x-\mu)^2\, p(x)$.

Standard Deviation: The standard deviation of X is the square root of the variance of X, denoted $\sigma = SD[X] = \sqrt{Var[X]}$.

The variance is the expected value of the squared deviations of the values of X from its mean μ.

PROPERTIES

Let X be a random variable with expected value μ.

1. $Var[X] = E[X^2] - E[X]^2 = E[X^2] - \mu^2$.
2. $Var[aX + c] = a^2\, Var[X]$ for any constants $a, c \in \mathbb{R}$.
3. $SD[aX + c] = |a|\, SD[X] = |a|\, \sigma$ for any constants $a, c \in \mathbb{R}$.

EXAMPLE

Recall the random variable X from a single roll of a fair die.

Then

$Var[X] = E[X^2] - \mu^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12} \approx 2.91667$ and

$\sigma = SD[X] = \sqrt{\frac{35}{12}} \approx 1.70783.$

APPLICATION: FINANCIAL DISTRESS AND AGENCY COST

INTRODUCTION

Assume company XYZ has a loan of $10 million that is due at the end of the year. Company XYZ is in financial distress because the market value of its assets will at the end of the year only be $9 million. In that case, company XYZ has to default on its debt. Company XYZ considers a new strategy with no upfront investment.

- **The probability of success of the new strategy is only 20%. Hence, the probability of failure of the new strategy is 80%.**
- **If the new strategy is successful, the value of the firm's assets will increase to $15 million.**
- **If the new strategy is not successful, the value of the firm's assets will fall to $5 million.**

	Old Strategy [mln]	Success [mln]	Failure [mln]
Value of Assets	9	15	5
Debt	9	10	5
Equity	0	5	0

QUESTION

Consider the table above. Calculate the expected value of the XYZ's assets under the new strategy. Is it beneficial for the equity and/ or bond holders if company XYZ executes this strategy?

Solution

We denote by the random variable X the possible values of XYZ's assets. We get:

- $E[X] = 0.2 \cdot \$15mln. + 0.8 \cdot \$5mln. = \$7mln.$ **Hence, under the new strategy, the expected value of XYZ's assets w**ill decrease **from** $\$9mln.$ **to** $\$7mln.$
- **If company XYZ does nothing, it will ultimately default and equity holders will get nothing with certainty. If the new strategy is implemented and succeeds, equity holders will get** $\$5$ **million in total. The expected payoff under the new strategy is equal to** $0.8 \cdot \$0mln. + 0.2 \cdot \$5mln. = \$1mln.$ **Therefore equity holders have**

nothing to lose **from this strategy.**

- **If company XYZ does nothing, it will ultimately default and debt holders will get $9 million with certainty. If the new strategy is implemented and succeeds, debt holders will get $10 million in total. If the new strategy is implemented and fails, debt holders will get $5 million in total. The expected payoff under the new strategy is equal to** $0.8 \cdot \$5mln. + 0.2 \cdot \$10mln. = \$6mln$. **Therefore debt holders have** ***a lot to lose*** **from this strategy.**

The results are summarized in the table below.

	Old Strategy [mln]	Success [mln]	Failure [mln]	Expected [mln]
Value of Assets	9	15	5	7
Debt	9	10	5	6
Equity	0	5	0	1

Effectively, the equity holders are gambling with the debt holder's money. Shareholders have an incentive to invest in negative NPV projects (where NPV stands for Net Present Value) that are risky, even though a negative NPV project may likely destroy value for the firm overall.

GENERAL DISTRIBUTIONS

CONTINOUS PROBABILITY SPACES

INTRODUCTION

Although sufficient for the development of many interesting topics in mathematical probability, the theory of discrete probability spaces does not go far enough for the rigorous treatment of problems of two kinds: those involving an infinitely repeated operation, as an infinite sequence of tosses of a coin, and those involving an infinitely fine operation, as the random drawing of a point from a segment. A mathematically complete development of probability, based on the theory of measure, puts these two classes of problems on the same footing, [...]. (Billingsley, Probability and Measure, Second Edition, 1986, page 1).

σ-ALGEBRA

MOTIVATION

In discrete probability, probabilities are defined for all possible events, that is, for all elements $E \in \mathcal{F}$, where $\mathcal{F}$ denotes the set of all subsets ($\mathcal{F} = 2^{\Omega}$) of the finite or countably infinite sample space Ω. Put differently, every subset of Ω has a well-defined probability.

This property is no longer true for continuous sample space with uncountably many elements. Intuitively, a continuous sample space may have such “crazy” subsets that we cannot assign them a proper probability. [Aside: No explicit representation of such sets exists, but their existence has been proven via the Axiom of Choice.]

As a result, we need to replace the power set of Ω by a smaller collection of subsets of Ω. This smaller collection of subsets must have certain properties so that we can define probabilities for its elements.

DEFINITION

σ-Algebra: Let Ω be an arbitrary nonempty space. A σ-algebra of Ω is a collection $\mathcal{F}$ of subsets of Ω such that

(1) $\Omega \in \mathcal{F}$;

(2) for any $E \in \mathcal{F}$ also $E^c \in \mathcal{F}$;

(3) For any countable subcollection $\{E_i\}_{i \in I}$ with $E_i \in \mathcal{F}$ for all $i \in I$, also $\bigcup_{i \in I} E_i \in \mathcal{F}$.

CONSEQUENCES

Statement (1) guarantees that the set $\mathcal{F}$ is nonempty. Conditions (2) and (3) state that $\mathcal{F}$ is closed under complements and countable unions, respectively. Since $\emptyset$ and Ω are complements, statement (1) is in the presence of condition (2) equivalent to $\emptyset \in \mathcal{F}$.

CONDITION

Recall de Morgan's laws, $(\bigcup_{i \in I} E_i)^c = \bigcap_{i \in I} E_i^c$ and $(\bigcap_{i \in I} E_i)^c = \bigcup_{i \in I} E_i^c$ for any finite or countably infinite index set I.

So, if $\mathcal{F}$ is closed under complements, then it is closed under countable unions if and only if it is closed under countable intersections. Therefore, we could replace condition (3) by the following condition:

(3') For any countable sub-collection $\{E_i\}_{i \in I}$ with $E_i \in \mathcal{F}$ for all $i \in I$, also $\bigcap_{i \in I} E_i \in \mathcal{F}$.

EXAMPLE

Examples of σ – algebra of a nonempty space Ω:

1. $F = \{\emptyset, \Omega\}$;
2. $F = \{\emptyset, A, A^c, \Omega\}$; for a set $\emptyset \neq A \subset \Omega$ with $A \neq \Omega$.
3. $\mathcal{F} = 2^{\Omega}$, the power set of Ω.

PROBABILITY MEASURE SPACE

We can define general probability spaces (including those with a continuous sample space).

DEFINITIONS

Probability Measure: Let Ω be an arbitrary nonempty sample space; let $\mathcal{F}$ be a σ – algebra of Ω. A function $P : \mathcal{F} \to \mathbb{R}$ is called a probability measure if it satisfies the following conditions:
(1) $0 \leq P(E) \leq 1$ for any event $E \in \mathcal{F}$.
(2) $P(\Omega) = 1$.
(3) Countable additivity: For any collection of pairwise disjoint events,
$E_i \in \mathcal{F}, \ i = 1, 2, \ldots,$
$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.

Probability Space: The triple $(\Omega, \mathcal{F}, P)$ is called a probability measure

space, or simply a probability space.

Events: The elements of the σ – algebra $\mathcal{F}$ are called events.

PROBABILITY RULES

The following properties are a simple generalization from discrete probability.

Let $(\Omega, \mathcal{F}, P)$ be a probability measure space. If $E, F \epsilon \mathcal{F}$, then

1. $P(E^c) = 1 - P(E)$;
2. If $E \subset F$, then $P(E) \leq P(F)$;
3. $P(E \bigcup F) = P(E) + P(F) - P(E \bigcap F)$.

The notions of

4. **Independence**
5. **Conditional probability**
6. **Bayes' rule**

naturally extend from discrete probability spaces to general probability measure spaces.

CONTINUOUS RANDOM VARIABLES

INTRODUCTION

Recall from discrete probability that the inverse image $X^{-1}(a)$ is always a well-defined event, that is, an element of the power set $\mathcal{F}$. Therefore, the probability mass function p assigning a probability $p(a)$ to any real number $a \in \mathbb{R}$ is (trivially) well-defined because every subset of the sample space has a well-defined probability.

This approach does not extend to general probability spaces. The probability of a singleton event is almost always zero; and not every subset of the sample space has a well-defined probability. Therefore, the definition of a random variable on a general probability space requires more care.

RANDOM VARIABLE

DEFINITION

Random Variable: Let $(\Omega, \mathcal{F}, P)$ be a probability measure space. A random variable is a function $X : \Omega \to \mathbb{R}$, such that for all $a \in \mathbb{R}$ the set $X^{-1}(-\infty, a]$ is in $\mathcal{F}$: $\{\omega \in \Omega : X(\omega) \leqslant a\} \in \mathcal{F}$.

PROPOSITION

Let X be a random variable on a probability measure space $(\Omega, \mathcal{F}, P)$. Then for all $a \in \mathbb{R}$ the following sets are in $\mathcal{F}$:

1. $X^{-1}(-\infty, a) = \{\omega \in \Omega : X(\omega) < a\}$,
2. $X^{-1}[a, \infty) = \{\omega \in \Omega : X(\omega) \geqslant a\}$,
3. $X^{-1}(a, \infty] = \{\omega \in \Omega : X(\omega) > a\}$,
4. $X^{-1}(a) = \{\omega \in \Omega : X(\omega) = a\}$.

DEFINITIONS

Let X be a random variable on a probability measure space $(\Omega, \mathcal{F}, P)$.

Cumulative Density Function (CDF): The CDF of X is the function

$F : \mathbb{R} \to [0, 1]$, such that for all $a \in \mathbb{R}$, $F(a) = P\left(X^{-1}(-\infty, a]\right)$.

Probability Density Function (PDF): An integrable function $f : \mathbb{R} \to \mathbb{R}_+$ is called the PDF of X if for all $a, b \in \mathbb{R}$ with $a \leq b$ it holds that $F(b) - F(a) = P\left(X^{-1}(a, b]\right) = \int_a^b f(x)\, dx$.

Let X be a random variable on the probability measure space $(\Omega, \mathcal{F}, P)$ with probability density function f.

Expected Value: The expected value of X, denoted $\mu = E[X]$, is given by $\mu = E[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx$.

Variance: The variance of X, denoted $\sigma^2 = Var[X] = E\left[(X - \mu)^2\right]$, is given by $Var[X] = E\left[(X - \mu)^2\right] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$.

PROBABILITY SPACES WITH Ω⊂ℝ

DEFINITION

Borel σ-Algebra: Let $\mathcal{A} = \{(-\infty, a];\ a \in \mathbb{R}\}$. Define $\mathcal{B} = \bigcap\{\mathcal{F} : \mathcal{F}$ is a σ –algebra such that $\mathcal{F} \supset \mathcal{A}\}$. We call $\mathcal{B}$ the σ – algebra generated by all intervals; it is the smallest σ – algebra which contains all intervals and is called the Borel σ-algebra.

Borel Sets: The elements of $\mathcal{B}$ are called Borel sets.
Any probability measure $P : \mathcal{B} \to \mathbb{R}$ leads to a well-defined probability measure space $(\mathbb{R}, \mathcal{B}, P)$.

EXAMPLE

Consider the random variable X with the continuous uniform distribution on the interval $[a, b] \subset \mathbb{R}$; then $f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \in \mathbb{R} \setminus [a, b] \end{cases}$

The random variable X has mean $\mu = (a + b)/2$ and variance $\sigma^2 = \frac{(b-a)^2}{12}$.

```
μ = Simplify[Integrate[x * 1 / (b - a), {x, a, b}]]
Simplify[Integrate[(x - μ)^2 * 1 / (b - a), {x, a, b}]]
```

$$\frac{a+b}{2}$$

$$\frac{1}{12}(a-b)^2$$

COVARIANCE AND CORRELATION

SUMS OF RANDOM VARIABLES

PROPERTIES

Properties: Let X, Y be random variables with expected values μ_X, μ_Y, and variances σ_X^2, σ_Y^2, respectively. For any constants a, b, $c \in \mathbb{R}$, :

- $E[aX + bY + c] = a\,E[X] + b\,E[Y] + c$ and
- $Var[aX + bY + c] = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab(E[XY] - \mu_X\mu_Y)$.

Note that we do not impose any restrictions on the random variables X and Y. They may be discrete, continuous, or even "mixed" (that is, they may have mass points (like a discrete random variable) as well as 'continuous parts'.)

COVARIANCE AND CORRELATION

DEFINITIONS

Covariance: Let X, Y be random variables with expected values μ_X, μ_Y, respectively. Then the covariance of X and Y is defined as
$Cov[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X\mu_Y$.

Correlation (Coefficient): If X, Y have positive variances σ_X^2, σ_Y^2, respectively, then the correlation ρ of X and Y is defined as
$\rho = \rho[X, Y] = \frac{Cov[X,Y]}{\sigma_X\sigma_Y}$.

PROPERTIES

Let X, Y be independent random variables with expected values μ_X, μ_Y, and variances σ_X^2, σ_Y^2, respectively. Then

- $Cov[X, Y] = E[XY] - \mu_X\mu_Y = 0$.
- For any constants $a, b, c \in \mathbb{R}$, $Cov[aX, bY] = 0$ and so,
- $Var[aX + bY + c] = a^2\sigma_X^2 + b^2\sigma_Y^2$.

We can easily extend the results to weighted finite sums of random variables.

PROPERTIES

Let $X_1, X_2, \ldots, X_n$ be random variables with expected values μ_i and variances σ_i^2, $i = 1, 2, \ldots, n$. For any constants $a_i, i = 1, 2, \ldots, n$, the random variable $W_n = \sum_{i=1}^{n} a_i X_i$ has

_ the mean $E[W_n] = \sum_{i=1}^{n} a_i \mu_i$

_ and the variance

$$Var[W_n] = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} a_i a_i Cov[X_i, X_j].$$

If the random variables $X_1, X_2, \ldots, X_n$ are mutually independent and have identical means $\mu_i \equiv \mu$ and variances $\sigma_i^2 = \sigma^2$, then the random variable $Y_n = \frac{\sum_{i=1}^{n} X_i}{n}$ has the mean $E[Y_n] = \mu$ and the variance $\frac{\sigma^2}{n}$.

LIMIT THEOREMS

INEQUALITIES

MARKOV INEQUALITY

DEFINITION

Markov Inequality: Suppose X is a nonnegative random variable. Then for any $a > 0$ it holds that $P(X \geq a) \leq \frac{E[X]}{a}$.

PROOF

We derive the inequality for continuous random variables with density f. For any $a > 0$ we can write:

$$E[X] = \int_0^\infty t\, f(t)\, dt = \int_0^a t\, f(t)\, dt + \int_a^\infty t\, f(t)\, dt \geq$$
$$\int_a^\infty t\, f(t)\, dt \geq \int_a^\infty a\, f(t)\, dt \geq a\, P(X \geq a)$$

CHEBYSHEV INEQUALITY

Next we state and prove the perhaps best-known inequality in probability theory.

DEFINITION

Chebyshev Inequality: Suppose X is a random variable with mean μ and variance σ^2. Then for any $k > 0$ it holds that $P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$. Equivalently, it holds that $P(|X - \mu| \geq n\sigma) \leq \frac{1}{n^2}$, for any $n > 0$.

PROOF

By the Markov inequality, $P\left((X-\mu)^2 \geq k^2\right) \leq \frac{\sigma^2}{k^2}$, since $(X-\mu)^2$ is a nonnegative random variable and $\sigma^2 = E\left[(X-\mu)^2\right]$. Observe that for any random variable X it holds that $|X - \mu| \geq k \iff (X-\mu)^2 \geq k^2$. Therefore, $P(|X - \mu| \geq k) = P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$. Setting $k = n\sigma$ we obtain the second version of the Chebyshev Inequality.

We use the Chebyshev Inequality to prove the Weak Law of Large Numbers.

WEAK LAW OF LARGE NUMBERS

BASICS

A sequence of random variables $X_1, X_2, \ldots, X_n$ is called independent and identically distributed, or i.i.d. for short, if the n random variables are independent and have all the same probability distribution. The latter assumption implies that they have identical mean μ and identical variance σ^2.

DEFINITION

Weak Law of large Numbers: Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Then for any $\epsilon > 0$ it holds that $P\left(\left|\frac{X_1+X_2+\cdots+X_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$. In particular, as $n \to \infty$, we have $P\left(\left|\frac{X_1+X_2+\cdots+X_n}{n} - \mu\right| \geq \epsilon\right) \to 0$.

In words: The probability that the average of the first n terms of a sequence of independent and identically distributed random variables differs from its mean by more than $\epsilon > 0$ tends to zero as n goes to infinity.

PROOF

Recall that the random variable $Y_n = \frac{\sum_{i=1}^{n} X_i}{n}$ has the mean $E[Y_n] = \mu$ and the variance $\frac{\sigma^2}{n}$. The Chebyshev Inequality implies $P(|Y_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$. And since $\frac{\sigma^2}{n\epsilon^2} \to 0$ for $n \to \infty$ the limit result follows.

Observe that the Weak Law of Large Numbers states that a particular probability involving a sequence of random variables tends to zero. This mode of convergence is an example of a general convergence concept in probability theory called convergence in probability.

DEFINITION

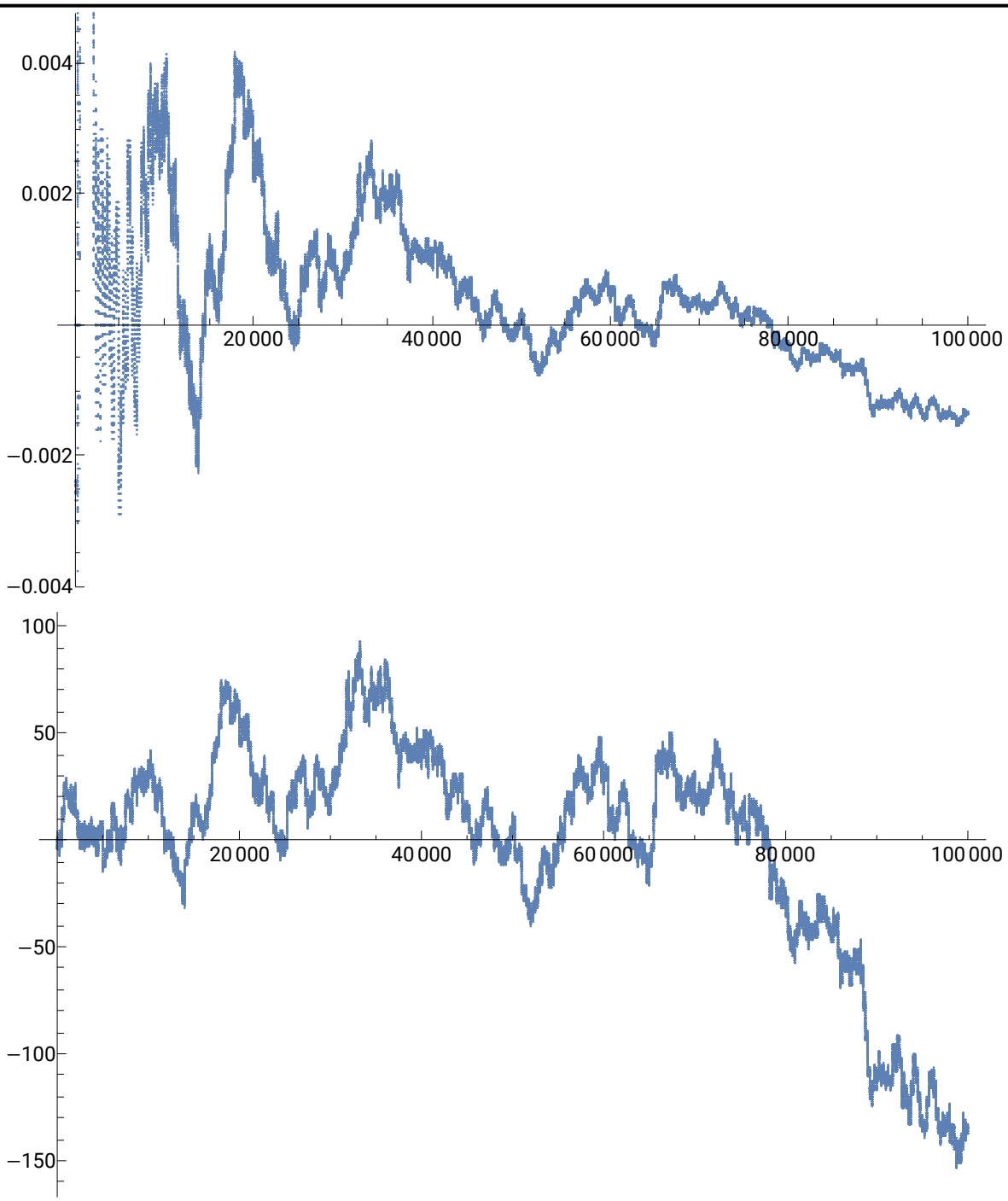
Convergence: A sequence $X_1, X_2, \ldots, X_n, \ldots$ of random variables converges to X in probability if for each $\epsilon > 0$ it holds that $P(|X_n - X| \geq \epsilon) \longrightarrow 0$ as $n \longrightarrow \infty$. Convergence in probability is denoted by $X_n \xrightarrow{P} X$ or $plim_{n\longrightarrow\infty} X_n = X$.

We illustrate the law with a Monte Carlo simulation example.

EXAMPLE TWO-STATE ECONOMY

Consider an economy that can be in two states, “bad” (0) or “good” (1). The random variable $X_i \in \{0, 1\}$ describes the state of the economy in period

$i = 1, 2, \ldots, n, \ldots$. Then the random variable $Y_n = \frac{\sum_{i=1}^{n} X_i}{n}$ denotes the proportion of time periods in which the economy is in the good state. Suppose $P(X_i = 0) = P(X_i = 1) = 0.5$. We simulate the state of the economy over n time periods.



The i.i.d. random variables $X_i, i = 1, 2, \ldots, n, \ldots$, have mean $\mu = \frac{1}{2}$ and variance $\sigma^2 = \frac{1}{4}$. The Weak Law of Large Numbers implies for any $\epsilon > 0$ that $P\left(\left| \frac{X_1 + X_2 + \cdots + X_n}{n} - \frac{1}{2} \right| \geq \epsilon \right) \leq \frac{1}{4n\epsilon^2}$. For example, for $n = 10\,000$ and

$\epsilon = 0.01$ the bound on the right-hand side is $\frac{1}{4\,n\epsilon^2} = \frac{1}{4} = 25\ \%$.

The law does **NOT** say that $\sum_{i=1}^{n} X_i \to n\mu$ as $n \to \infty$. The Chebyshev Inequality only implies that $P\left(\left| \sum_{i=1}^{n} X_i - \frac{1}{2}\, n \right| \geqslant n\epsilon \right) \leqslant \frac{1}{4\,\epsilon^2}$ since the variance of $Var[\sum_{i=1}^{n} X_i] = \frac{1}{4}\, n$. In this case, the bound on the right-hand side is independent of n. (Aside: In fact, $\left| \sum_{i=1}^{n} X_i - \frac{1}{2}\, n \right| \longrightarrow \infty$ with probability 1.)

STRONG LAW OF LARGE NUMBERS

DEFINITION

Strong Law of large Numbers: Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. random variables with mean μ. Then, with probability 1, that $\frac{X_1+X_2+\cdots+X_n}{n} \longrightarrow \mu$ as $n \longrightarrow \infty$; that is, $P\left(lim_{n\to\infty} \frac{X_1+X_2+\cdots+X_n}{n} = \mu\right) = 1.$

The convergence result in the Strong Law of Large Numbers is also an example of a general convergence concept in probability theory, namely of almost sure convergence. This mode of convergence in probability theory comes closest to pointwise convergence of functions in real analysis.

DEFINITION

Almost sure Convergence: A sequence $X_1, X_2, \ldots, X_n, \ldots$ of random variables converges almost surely or with probability 1 to X if it holds that $X_n \longrightarrow X$ with probability 1 as $n \longrightarrow \infty$; that is, $P\left(lim_{n\longrightarrow\infty} X_n = X\right) = 1.$

Almost sure convergence is denoted by $X_n \overset{a.s.}{\longrightarrow} X$.

CENTRAL LIMIT THEOREM

BASICS

The most famous theorem in probability theory is the Central Limit Theorem. (We omit its nontrivial proof.)

DEFINITION

Central Limit Theorem: Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Let $S_n = \sum_{i=1}^{n} X_i$. Then the distribution of the random variable $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ tends to the standard normal distribution as $n \to \infty$. That is,

$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq y\right) \to \Phi(y)$ with $\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-x^2/2}\, dx$ as $n \to \infty$.

Observe that the Central Limit Theorem states that the distributions of a sequence of random variables tend to the probability distribution of a particular random variable. This mode of convergence is an example of a general convergence concept called convergence in distribution or weak convergence.

DEFINITION

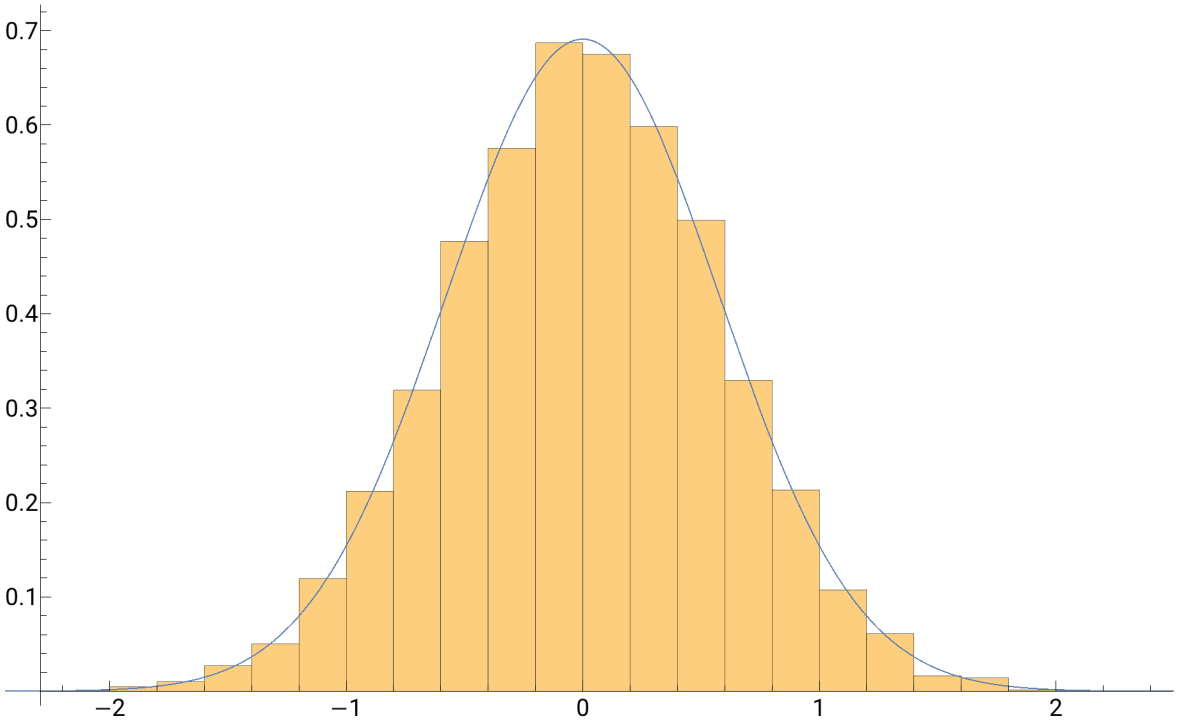
Weak Convergence: Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of random variables with cumulative distribution functions $F_1, F_2, \ldots, F_n$. Also let X be a random variable with cdf F. Then the sequence $X_1, X_2, \ldots, X_n, \ldots$ is said to converge in probability, or to converge weakly, to X if $lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}$ at which F is continuous. Convergence in probability is denoted by $X_n \xrightarrow{\mathcal{D}} X$ or $X_n \Longrightarrow X$.

The Central Limit Theorem is of fundamental importance to statistics. Note that $\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{Y_n - \mu}{\sigma/\sqrt{n}}$ and so, in the language of statistics, the sampling distribution of the sample average $Y_n = \frac{\sum_{i=1}^{n} X_i}{n}$ of a large sample tends to the standard normal distribution.

SIMULATION

Consider a sequence of i.i.d. random variables with a uniform distribution on the interval $[-10, 10]$; these random variables have mean $\mu = 0$ and variance

$\sigma^2 = \int_{-10}^{10} x^2 \frac{1}{20} dx = \frac{100}{3}$. Then the average $Y_n = \frac{\sum_{i=1}^{n} X_i}{n}$ has mean 0 and variance $\frac{\sigma^2}{n}$. Now consider repeated simulations of the average Y_{100}.



{μ → 0.0050777, σ → 1.00109}

ONE MORE IMPORTANT INEQUALITY

CONVEX AND CONCAVE FUNCTIONS

Convex and concave functions have many applications in economics and finance.

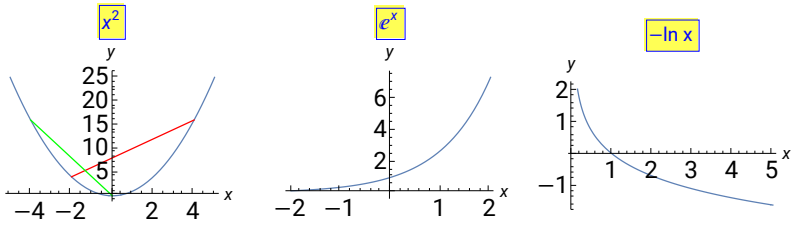
DEFINITION

Convex: Let $I \subset \mathbb{R}$ be an interval. A function $f : I \to \mathbb{R}$ is called convex if $f(\lambda\, x_1 + (1-\lambda)\, x_2) \leq \lambda\, f(x_1) + (1-\lambda)\, f(x_2)$ for all $\lambda \in [0,\ 1]$ and any $x_1,\ x_2 \in \mathbb{R}$.

Strictly Convex: If the above inequality is strict whenever $x_1 \neq x_2$ and $0 < \lambda < 1$, then f is called strictly convex.

ILLUSTRATION

The functions (with the terms) x^2, e^x, and $-ln(x)$ are convex.

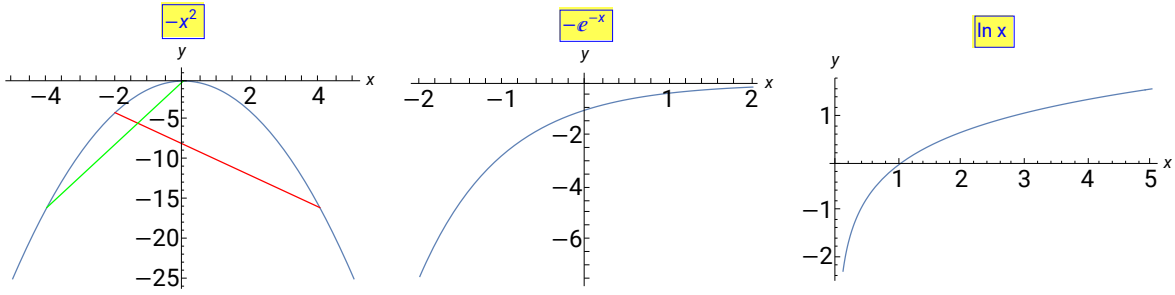


Concave: A function $f : I \to \mathbb{R}$ is called concave if $-f$ is convex; that is, if $f(\lambda\, x_1 + (1-\lambda)\, x_2) \geq \lambda\, f(x_1) + (1-\lambda)\, f(x_2)$ for all $\lambda \in [0,\ 1]$ and any $x_1,\ x_2 \in \mathbb{R}$.

Strictly Concave: If the above inequality is strict whenever $x_1 \neq x_2$ and $0 < \lambda < 1$, then f is called strictly concave.

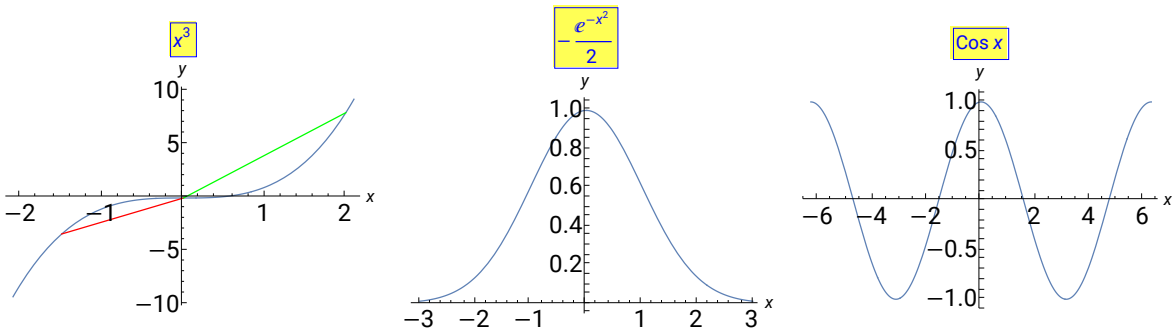
ILLUSTRATION

The functions (with the terms) $-x^2$, $-e^{-x}$, and $ln(x)$ are concave.



NEITHER CONVEX NOR CONCAVE FUNCTIONS

The functions (with the terms) x^3, $cos(x)$, and $ln(x)$ are neither convex nor concave. The function $f : (-\infty, 0] \to \mathbb{R}$ with $f(x) = x^3$ is concave on its domain $(-\infty, 0]$. The function $g : [0, \infty) \to \mathbb{R}$ with $g(x) = x^3$ is convex on its domain $[0, \infty)$.

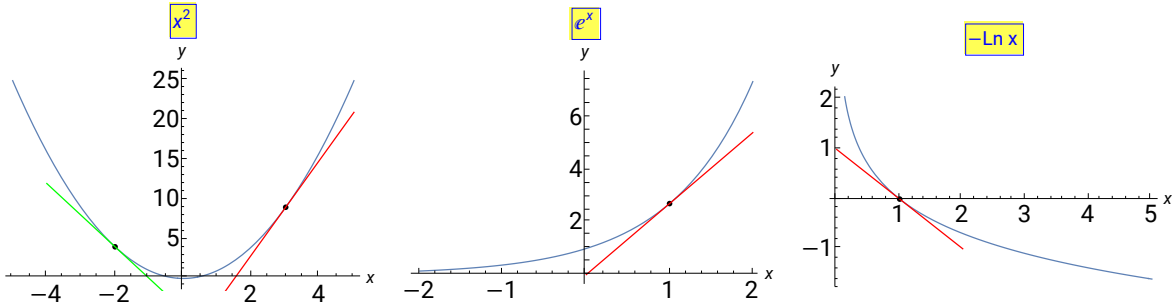


PROPERTIES OF CONVEX FUNCTIONS

Let $I \subset \mathbb{R}$ be an interval. A differentiable function $f : I \to \mathbb{R}$ is convex if and only if for any $y \in I$ $f(x) \geq f(y) + f'(y)(x - y)$ for all $x \in I$.

ILLUSTRATION

The graph of a convex function “lies” entirely above any tangent to the graph.

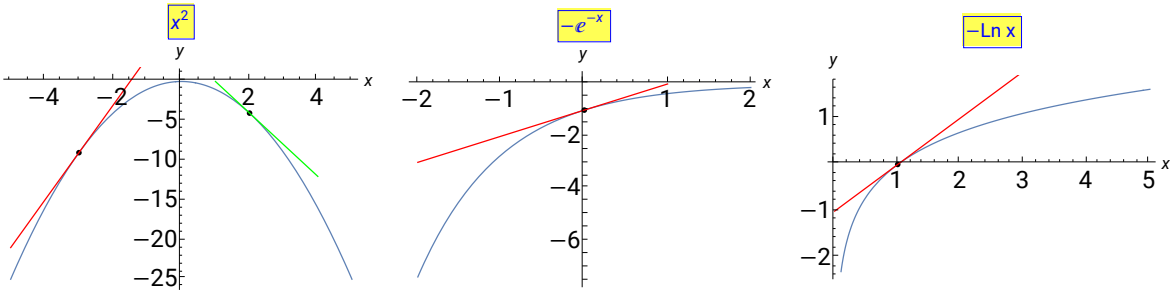


PROPERTIES OF CONCAVE FUNCTIONS

A differentiable function $f : I \to \mathbb{R}$ is concave if and only if for any $y \in I$ $f(x) \leqslant f(y) + f'(y)(x - y)$ for all $x \in I$.

ILLUSTRATION

The graph of a convex function "lies" entirely above any tangent to the graph.



JENSEN'S INEQUALITY

DEFINITION

Jensen's Inequality: Let X be a random variable (with finite mean) and $f : \mathbb{R} \to \mathbb{R}$ a differentiable convex function. Then $E\left[f(X)\right] \geqslant f(E[X])$. If $g : \mathbb{R} \to \mathbb{R}$ is a differentiable concave function, then $E\left[g(X)\right] \leqslant g(E[X])$.

PROOF

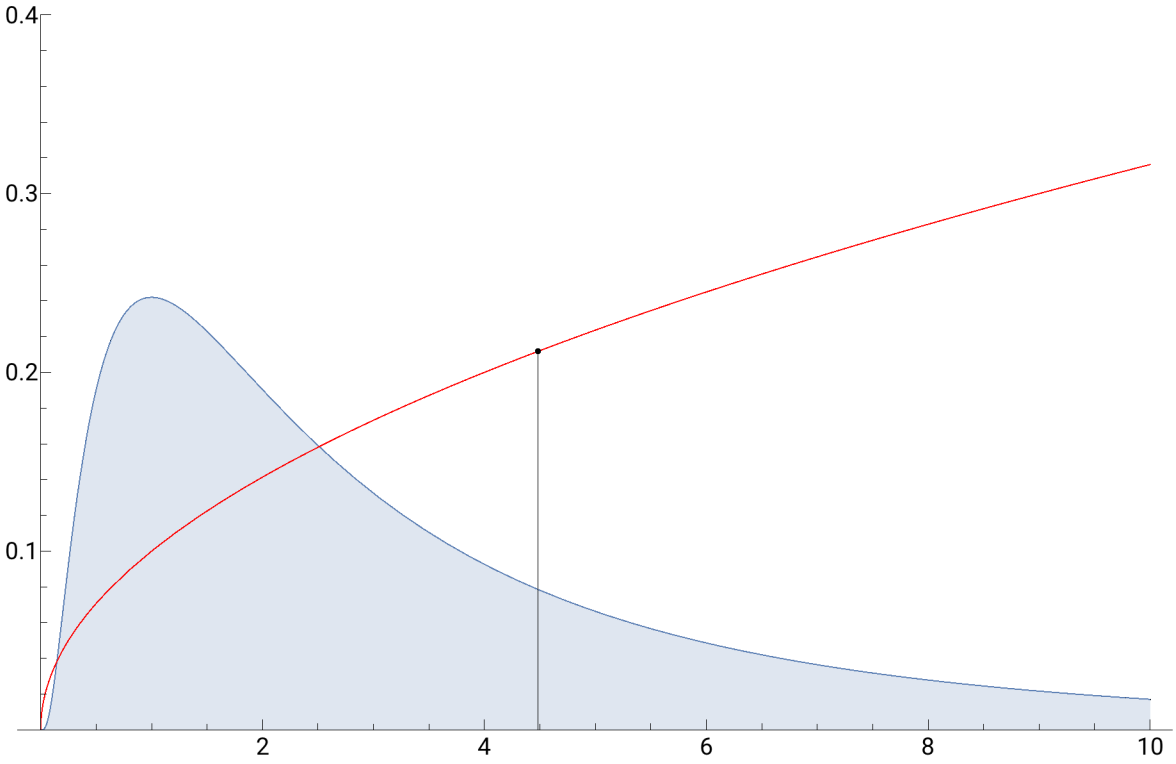
Let $\mu = E[X]$. Since $f : \mathbb{R} \to \mathbb{R}$ is differentiable and convex, $f(x) \geqslant f(\mu) + f'(\mu)(x - \mu)$ for all $x \in \mathbb{R}$ and so $f(X) \geqslant f(\mu) + f'(\mu)(X - \mu)$ for the random variable X. Taking expectations on both sides of the inequality yields $E\left[f(X)\right] \geqslant f(\mu) + f'(\mu)\,E[X - \mu]\big] = E[\mu]$.

APPLICATION IN ECONOMICS

Consider two investments. Investment 1 has a random return described by a random variable X with mean $m = E[X]$. Investment 2 is a risk-free investment with a sure return of m. An investor evaluates the investments using a utility function $u : \mathbb{R} \to \mathbb{R}$. If the investor is risk-averse then her utility function u is concave; thus $E[u(X)] \leqslant u(E[X]) = u(m)$ and the investor prefers the risk-free investment 2. On the contrary, if the investor is risk-seeking then her utility function u is convex; thus $E[u(X)] \geqslant u(E[X]) = u(m)$ and the investor prefers the risky investment 1.

NUMERICAL EXAMPLE

Suppose the return X of investment 1 has a log-normal probability distribution with parameters $\mu = 1$ and $\sigma = 1$. Then the expected return is $m = E[X] = e^{\mu + \frac{\sigma^2}{2}} \approx 4.4817$. The risk-free investment 2 has return m. The investor's utility function is $u(x) = \frac{\sqrt{x}}{10}$.




```
0.186825

0.2117
```

The investor's expected utility from investment 1 is $E[u(X)] \approx 0.1868$. Her utility from the risk-free investment 2 is $u(m) = e^{\frac{3}{4}} \approx 0.2117$.