Probability

Logit

2016

Table of Contents

- Discrete Math (Factorial, Combinations, Set Theory)
- Probability Basics (Event, Sample Space, etc.)
- Axioms of Probability
- Dependent vs. Independent Random Variables
- Conditional Probability
- Bayes' Theorem
- Continuous vs. Discrete Random Variable
- CDF, PMF, PDF
- Expected Value, Variance, Standard Deviation
- Covariance, Correlation
- Discrete Distributions (Bernoulli, Binomial, Poisson, Geometric)
- Continuous Distributions (Uniform, Normal, Exponential, χ^2)

Intro to Probability

Probability is the study of random events. Probability theory tries to quantify uncertainty and organize the unpredictable. An understanding of probability is fundamental for statistics and machine learning.

Intro to Probability

To do probability, we first need to be able to *count* all possibilities for a given problem. To deal with *events* and *event spaces*, we will have to review some fundamentals of combinatorics and set theory from discrete mathematics.

Discrete Math Fundamentals

Discrete Math Fundamentals

At a basic level, counting is aided by an understanding of combinations, permutations, and factorials.

Combinations is the number of arrangements of objects where order *doesn't matter*.

Permutations is the number of arrangements of objects where order *does matter*.

Factorial is the number of permutations of a set of objects without repetition.

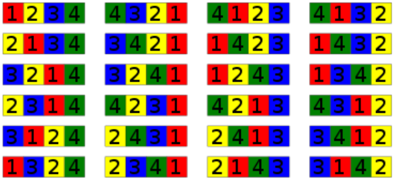


Image Source

All permutations of a set of 4 distinct objects.

Discrete Math Fundamentals

Factorial:

$$n! = (n) \cdot (n-1) \cdots (2) \cdot 1$$

Permutations of r elements from n possible choices, *with repetition*:

$$n^r$$

Permutations of r elements from n possible choices, *without repetition*:

$${}_nP_r = \frac{n!}{(n-r)!}$$

Combinations of r elements from n possible choices, *with repetition*:

$$\binom{n+r-1}{r} = \binom{n+r-1}{n-1} = \frac{(n+r-1)!}{r!(n-1)!}$$

Combinations of r elements from n possible choices, *without repetition*:

$${}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Discrete Math Fundamentals

How many ways can we create a 5-letter password using each of the letters $\{A, B, C, D, E\}$ only once?

Discrete Math Fundamentals

How many ways can we create a 5-letter password using each of the letters $\{A, B, C, D, E\}$ only once?

We have 5 possibilities for the first letter, then 4 possibilities for the second letter, and so on until the last letter.

$$\underline{5} \cdot \underline{4} \cdot \underline{3} \cdot \underline{2} \cdot \underline{1} = 5! = 120$$

There are 5! (5 factorial), or 120, arrangements of these 5 letters.

Discrete Math Fundamentals

How many ways can we choose 5 different letters from the alphabet?

Discrete Math Fundamentals

How many ways can we choose 5 different letters from the alphabet?

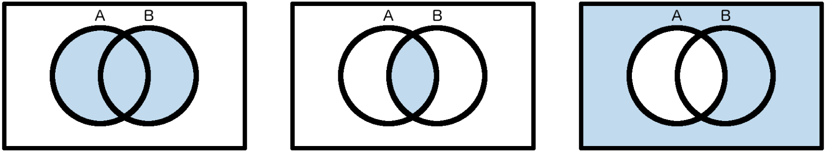
We have 26 choices for the first letter, 25 choices for the second letter, etc. Note that we don't care about order, so we divide by the number of ways we can order those letters.

$$\binom{26}{5} = \frac{26!}{5!(26-5)!} = \frac{26 \cdot 25 \cdot 24 \cdot 23 \cdot 22}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 65780$$

In total, this is $\binom{26}{5}$ (pronounced "26 choose 5"), or 65780, possibilities.

Discrete Math Fundamentals

Set theory notation:

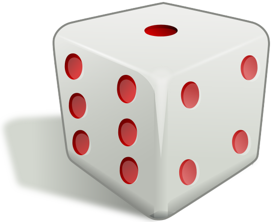


Left: $A \cup B$: the union of A and B.
Center: $A \cap B$: the intersection of A and B.
Right: A^C: the complement of A.

Probability

Example: Die Roll

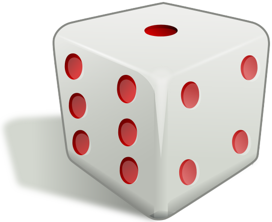
Suppose we roll a fair 6-sided die once. What is the probability that we will roll either a 2 or a 3?



Example: Die Roll

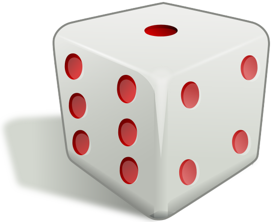
Suppose we roll a fair 6-sided die once. What is the probability that we will roll either a 2 or a 3?

Assuming the probability for rolling each number is equal, then the probability for rolling any given number is $\frac{1}{6}$. Consequently, the probability of rolling either 2 or 3 is $\frac{1}{3}$.



Example: Die Roll

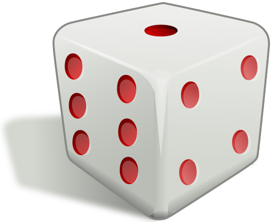
What is the probability that we roll *neither* 2 *nor* 3?



Example: Die Roll

What is the probability that we roll *neither* 2 *nor* 3?

The probability of rolling either 2 or 3 is $\frac{1}{3}$, and since the probability of all disjoint events must add up to 1, the probability of *not* rolling 2 or 3 is $\frac{2}{3}$.



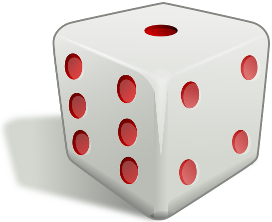
Sample Space

In probability, an event is a set of outcomes of an experiment. A sample space S is the set of all possible outcomes. In the case of a die roll, the sample space is

$$S = \{1, 2, 3, 4, 5, 6\}$$

We are completely certain that one of the events from the sample space will occur. In more mathematical terms, with the probability function $\mathbb{P}$,

$$\mathbb{P}(S) = 1$$



Events

If we set X and Y to be the events:

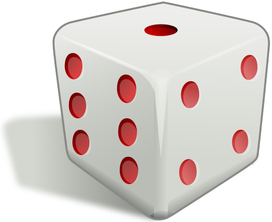
$$X = \text{ roll a 2}$$
$$Y = \text{ roll a 3}$$

Then X and Y are disjoint events, since $X \cap Y = \varnothing$ (empty set). For any two disjoint events X and Y,

$$\mathbb{P}(X \cup Y) = \mathbb{P}(X) + \mathbb{P}(Y)$$

Since the probability of rolling any given number is $\frac{1}{6}$, we know

$$\mathbb{P}(X \cup Y) = \mathbb{P}(X) + \mathbb{P}(Y) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$



Events

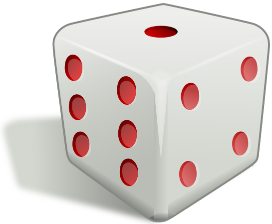
If we set X, Y, and Z to be the events:

$$\begin{aligned} X &= \text{ roll a 2} \\ Y &= \text{ roll a 3} \\ Z = X \cup Y &= \text{ roll a 2 or 3} \\ Z^C &= \text{ roll neither 2 nor 3} \end{aligned}$$

We know that the probability for all disjoint events in S have to add up to 1. This means

$$\mathbb{P}(S) = \mathbb{P}(Z \cup Z^C) = \mathbb{P}(Z) + \mathbb{P}(Z^C) = 1$$

By this logic, since $\mathbb{P}(Z) = \frac{1}{3}$, then $\mathbb{P}(Z^C) = \frac{2}{3}$.



Axioms of Probability

A probability function $\mathbb{P}$ has to abide by the Axioms of Probability:

- $0 \leq \mathbb{P}(A) \leq 1$ for each event A.
- $\mathbb{P}(S) = 1$ for the entire space S (the "sure event").
- For a sequence $A_1, A_2, \ldots$ of mutually disjoint events,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Example: Coin Flip (Independence)

Amy and Beth each flip a coin once. Assuming both coins are fair, what is the probability that both Amy and Beth get tails?

Example: Coin Flip (Independence)

Amy and Beth each flip a coin once. Assuming both coins are fair, what is the probability that both Amy and Beth get tails?

Define the following events:

$$A_H = \text{Amy flips heads}, \quad A_T = \text{Amy flips tails},$$
$$B_H = \text{Beth flips heads}, \quad B_T = \text{Beth flips tails}$$

We assume Amy's and Beth's coin flips are independent events, so we can multiply their respective probabilities together:

$$\begin{aligned} \mathbb{P}(A_T \cap B_T) &= \mathbb{P}(A_T) \cdot \mathbb{P}(B_T) \\ &= \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \end{aligned}$$

Example: Golf (Dependence)

Suppose Jerry likes to go golfing on weekends. Let G be the event Jerry goes golfing this weekend, and R the event that it's raining this weekend.

$$\mathbb{P}(G) = 0.30, \quad \mathbb{P}(R) = 0.20$$

If these were independent events, we'd expect the probability that Jerry goes golfing *and* it's raining this weekend to be

$$\mathbb{P}(G \cap R) = \mathbb{P}(G) \cdot \mathbb{P}(R) = 0.06$$

when, in reality, the rain may dissuade Jerry from golfing, making the probability closer to 1%. Because of this, we can say that these events are <u>dependent</u>.

Example: Golf (Conditional Probability)

Conditional probability gives us the probability of one event happening *given* that another event occurs.

The conditional probability $\mathbb{P}(G|R)$ that Jerry goes golfing *given* that it's raining can be calculated like so:

$$\begin{aligned}\mathbb{P}(G|R) &= \frac{\mathbb{P}(G \cap R)}{\mathbb{P}(R)} \\ &= \frac{0.01}{0.20} = 0.05\end{aligned}$$

If it's raining, there's only a 5% chance Jerry will go golfing this weekend.

Conditional Probability and Independence

We can say that event A is independent of event B if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

In the context of conditional probability, A is independent of B if

$$\mathbb{P}(A|B) = \mathbb{P}(A)$$

In other words, the probability of A given B is the same as the probability A occurs in general. If B occurs, it has no bearing on A.

Bayes' Theorem

If we start with the definition of conditional probability and do some algebra:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

$$\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}}$$

This last boxed equation is known as Bayes' theorem.

Example: Golf (Bayes' Theorem)

Using Bayes' theorem, what is the probability that it rains this weekend *given* that Jerry goes golfing? We know

$$\mathbb{P}(G) = 0.30, \quad \mathbb{P}(R) = 0.20,$$
$$\mathbb{P}(G|R) = 0.05$$

Example: Golf (Bayes' Theorem)

Using Bayes' theorem, what is the probability that it rains this weekend *given* that Jerry goes golfing? We know

$$\mathbb{P}(G) = 0.30, \quad \mathbb{P}(R) = 0.20,$$
$$\mathbb{P}(G|R) = 0.05$$

Using Bayes' theorem:

$$\begin{aligned} \mathbb{P}(R|G) &= \frac{\mathbb{P}(G|R)\mathbb{P}(R)}{\mathbb{P}(G)} \\ &= \frac{(0.05)(0.20)}{0.30} = 0.0333\ldots \end{aligned}$$

Random Variable

A random variable is a function from the set of possible outcomes from a random experiment to some other set (usually the set of real numbers).

An example of a random variable is rolling two dice and giving the sum of the numbers. While there are 36 possible outcomes for the two dice, the sum can only take on the values $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.

Discrete Random Variable

More specifically, a dice roll is an example of a discrete random variable because it can only take on values from a discrete set (either finite or countably infinite).

Continuous Random Variable

A continuous random variable, by contrast, can take on an uncountably infinite number of values.

For instance, the length of a phone call can be considered a continuous random variable.

Cumulative Distribution Function

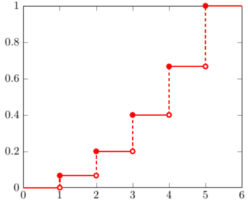


Image Source

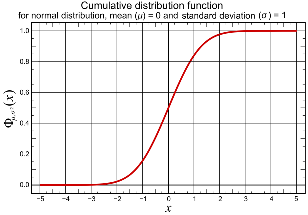


Image Source

A cumulative distribution function (CDF) gives us information about the distribution of a random variable, either discrete (on the left), or continuous (on the right). A CDF is defined as

$$F_X(x) = \mathbb{P}(X \leq x)$$

where X is our random variable, and x is some value.

Cumulative Distribution Function

A cumulative distribution function $F_X(x)$ has the following properties:

- $F_X(x)$ always takes values from 0 to 1 inclusive.
- $F_X(x)$ is monotonically increasing.
- $F_X(x)$ approaches 0 to the left, and approaches 1 to the right.
- $F_X(b) - F_X(a)$ gives us the probability that random variable X takes a value between a and b.

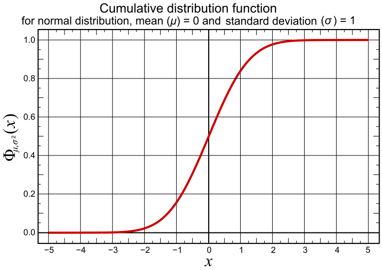


Image Source

Probability Mass Function (Discrete)

For discrete random variables, a probability mass function (PMF) gives us the probability that a random variable takes a specific value.

For instance if X represents the sum on a pair of dice, and $f_X(x)$ is its PMF, then

$$f_X(2) = \mathbb{P}(X = 2) = \frac{1}{36}$$
$$f_X(3) = \mathbb{P}(X = 3) = \frac{2}{36}$$
$$f_X(4) = \mathbb{P}(X = 4) = \frac{3}{36}$$
$$\vdots$$

Probability Density Function (Continuous)

For a *continuous* random variable, it no longer makes sense to talk about the probability of that variable taking a specific value. For instance, what is the probability that a given phone call is exactly 15 minutes long, and not a nanosecond longer or shorter? Virtually zero.

Instead, for continuous random variables, we use a probability density function (PDF).

Probability Density Function (Continuous)

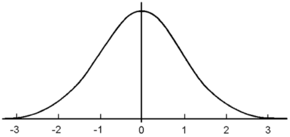


Image Source

A probability density function (PDF) $f_X(x)$ can give us information about the probability of X taking values in a range, rather than just a specific value.

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx$$

From this, we can give a relationship between the CDF and the PDF:

$$\mathbb{P}(X < x) = F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

Joint Probability Distribution

A joint probability distribution covers the cases where we want to investigate multiple random variables simultaneously.

In the discrete case for random variables X and Y, this is

$$\mathbb{P}(X = x, Y = y)$$

which is the probability that $X = x$ and $Y = y$ *simultaneously*.

In the continuous case for random variables X and Y, this involves joint probability density functions:

$$f_{X,Y}(x, y)$$

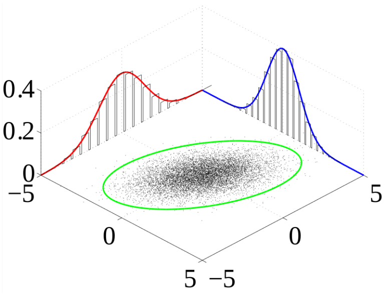


Image Source

Example: Golf (Joint Probability)

If

$$X = 0 \leftarrow \text{no golf} \qquad X = 1 \leftarrow \text{golf}$$
$$Y = 0 \leftarrow \text{no rain} \qquad Y = 1 \leftarrow \text{rain}$$

then we can describe the joint probability distribution like so:

	$X = 0$	$X = 1$
$Y = 0$	0.51	0.19
$Y = 1$	0.29	0.01

Example: Golf (Marginal Probability)

$X = 0 \leftarrow$ no golf $\qquad X = 1 \leftarrow$ golf

$Y = 0 \leftarrow$ no rain $\qquad Y = 1 \leftarrow$ rain

	$X = 0$	$X = 1$	SUM
$Y = 0$	0.51	0.19	0.70
$Y = 1$	0.29	0.01	0.30
SUM	0.80	0.20	1.00

We managed to calculate the remaining parts of the joint probability distribution by considering the marginal distribution - related to the sums along the margin of this table. We knew that the probability of golf was 0.30, the probability of rain was 0.20, and the probability of golfing on a rainy weekend was 0.01.

Expected Value

The expected value of a random variable is a kind of "weighted average" for all possible values that variable can take. The expected value is what we would tend to see on average after doing an experiment a large number of times.

Expected Value

For a discrete random variable X, the expected value is

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i p_i$$

where $\{x_i\}$ is the set of values X can take, and $\{p_i\}$ are the associated probabilities for those values.

For a continuous random variable Y, the expected value is

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f(y) dy$$

where $f(y)$ is the probability density function (PDF) for Y.

Expected Value Example: Die Roll

For instance, a die roll can take values in the set $\{1, 2, 3, 4, 5, 6\}$, and each value has probability $\frac{1}{6}$ of being rolled, so the expected value is

$$\mathbb{E}[X] = \sum_{i=1}^{6} x_i p_i$$
$$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6}$$
$$+ 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}$$

If you roll a die enough times, the average of the numbers rolled should approach $\frac{7}{2}$.

Variance and Standard Deviation

Variance is a measure of the "spread" of a probability distribution. A larger variance corresponds with a more spread-out distribution, with a greater proportion of the probability "weights" located further away from the mean. Standard deviation, another measure of spread, is just the square root of the variance.

Variance and Standard Deviation

In general, if μ is the mean (expected value) of a random variable X, the variance is defined as

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

For a discrete random variable X, this can also be defined as

$$\mathrm{Var}(X) = \sum_{i=1}^{n} p_i (x_i - \mu)^2$$

For a continuous random variable Y with density function $f(y)$,

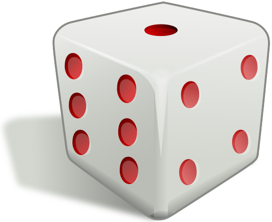
$$\mathrm{Var}(Y) = \int_{-\infty}^{\infty} (x-\mu)^2 f(y) dy$$

The standard deviance σ is just the square root of the variance, or equivalently,

$$\mathrm{Var}(X) = \sigma^2(X)$$

Variance and Standard Deviation Example

What is the variance and standard deviation for a fair die roll?

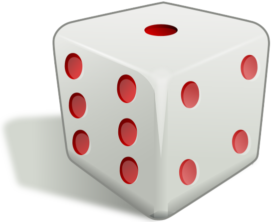


Variance and Standard Deviation Example

What is the variance and standard deviation for a fair die roll?

We know from before that $\mu = 3.5$.

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{i=1}^{6} p_i (x_i - \mu)^2 \\
&= \frac{1}{6}(1-3.5)^2 + \frac{1}{6}(2-3.5)^2 \\
&+ \frac{1}{6}(3-3.5)^2 + \frac{1}{6}(4-3.5)^2 \\
&+ \frac{1}{6}(5-3.5)^2 + \frac{1}{6}(6-3.5)^2 \\
&\boxed{\approx 2.917} \approx \sigma^2 \\
&\boxed{\sigma \approx 1.708}
\end{aligned}$$



Sum of Random Variables

For the sum of 2 random variables, X and Y, the mean or expected value is

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

and the variance is

$$\mathbb{V}\text{ar}(X+Y) = \mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(Y) + 2\text{Cov}(X,Y)$$

If X and Y are independent (not correlated), this simplifies to

$$\mathbb{V}\text{ar}(X+Y) = \mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(Y)$$

Covariance and Correlation

Covariance is a measure of the relationship between two random variables. For random variables X and Y, the covariance is:

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

A related measure of dependence between two random variables is correlation, which for random variables X and Y is defined as:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations for X and Y respectively. Correlation can only take values between -1 and 1 inclusive, and is dimensionless. Unlike covariance, correlation is not dependent on scaling.

Covariance Matrix

For a set of n random variables $\{X_1, \ldots, X_n\}$, the covariance matrix $\mathbf{\Sigma}$ is defined as

$$\mathbf{\Sigma} = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Cov}(X_n, X_n) \end{bmatrix}$$

Note that $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ and $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$, so this is a symmetric matrix (symmetric with respect to the diagonal) with the variances of the individual random variables on the diagonal.

Covariance Matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Cov}(X_n, X_n) \end{bmatrix}$$

The covariance matrix gives us a measure of how closely related each random variable is with respect to each other. We will see applications of the covariance matrix when doing *Principal Component Analysis (PCA)*, as the eigenvectors and eigenvalues of $\mathbf{\Sigma}$ will help determine the principal components.

Correlation

Correlation gives a dimensionless measure of the linear relationship between two random variables.

- Linear relationship, positive slope: correlation close to 1.
- Linear relationship, negative slope: correlation close to -1.
- Little evidence of linear relationship: correlation close to 0.

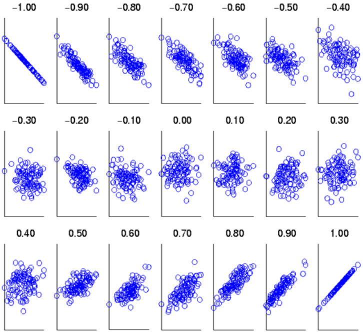


Image Source

Important Distributions

Important Distributions

- Discrete Distributions
  - Bernoulli
  - Binomial
  - Poisson
  - Geometric
- Continuous Distributions
  - Uniform
  - Normal / Gaussian
  - Exponential
  - χ^2

Bernoulli Distribution

The simplest distribution, a Bernoulli random variable can only take two values (e.g. 0 and 1). For example, a coin flip (not necessarily fair) is a Bernoulli random variable. Let p be the probability of "success" or "heads".

- Probability Mass Function: p for "heads", $(1 - p)$ for "tails".
- Mean: p.
- Variance: $p(1 - p)$.

Other examples of the Bernoulli distribution:

- Success of medical treatment.
- Student passing exam.
- Transmitting a disease.

Binomial Distribution

A binomial distribution describes the outcome from a series of Bernoulli experiments, e.g. a sequence of coin flips. This distribution approaches a Gaussian distribution in shape when the number of trials becomes large. Let n be the number of trials, and p be the probability of success.

- $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ (probability of k successes).
- Mean: np.
- Variance: $np(1-p)$.

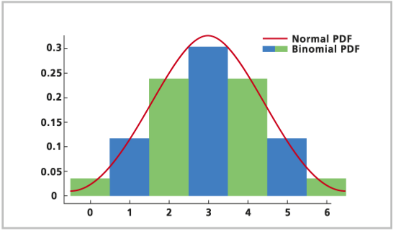


Image Source

Binomial Example (Basketball)

A basketball team has a 60% chance of winning any given game. What is the probability the team will win exactly 6 of the next 10 games?

Binomial Example (Basketball)

A basketball team has a 60% chance of winning any given game. What is the probability the team will win exactly 6 of the next 10 games?

$$\mathbb{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\begin{aligned}\mathbb{P}\{X = 6\} &= \binom{10}{6} (0.6)^6 (1-0.6)^{10-6} \\ &= 210(0.046656)(0.0256) \\ &\approx 0.2508\end{aligned}$$

Poisson Distribution

The Poisson distribution gives the probability of a certain number of events occurring in a fixed interval of time. The Poisson distribution is defined with a parameter λ.

- $\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$.
- Mean: λ.
- Variance: λ.

Examples of Poisson distributions include:

- Number of phone calls between 10-11 AM on any given weekday.
- Number of decay events per second from radioactive material.

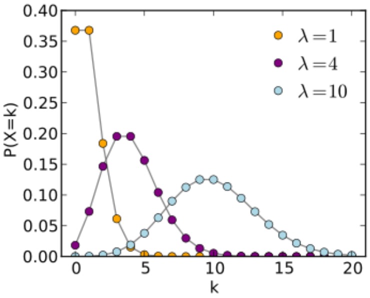


Image Source

Poisson Example (Call Center)

A call center averages 5 calls per hour. What is the probability it receives fewer than 3 calls in the next hour?

Poisson Example (Call Center)

A call center averages 5 calls per hour. What is the probability it receives fewer than 3 calls in the next hour?

We know $\lambda = 5$

$$\begin{aligned}
\mathbb{P}\{X = k\} &= \frac{\lambda^k e^{-\lambda}}{k!} \\
\mathbb{P}\{X < 3\} &= \mathbb{P}\{X = 0\} + \mathbb{P}\{X = 1\} \\
&\quad + \mathbb{P}\{X = 2\} \\
&= \frac{(5^0)(e^{-5})}{0!} + \frac{(5^1)(e^{-5})}{1!} \\
&\quad + \frac{(5^2)(e^{-5})}{2!} \\
&\approx 0.1247
\end{aligned}$$

Geometric Distribution

The geometric distribution describes the probability that a sequence of Bernoulli trials will have its first success on the k^{th} trial. For instance, what is the probability that a coin will come up only tails on the first $k-1$ trials, then finally come up heads on the k^{th} trail?

- $\mathbb{P}(X = k) = (1-p)^{k-1}p$.
- Mean: $\frac{1}{p}$.
- Variance: $\frac{1-p}{p^2}$.

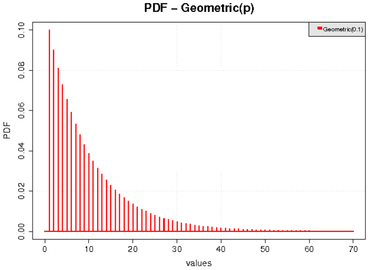
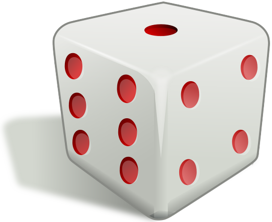


Image Source

Geometric Distribution Example (Die Roll)

What is the probability that we roll a fair 6-sided die 4 times before finally rolling a 5 (on the 5^{th} roll)?



Geometric Distribution Example (Die Roll)

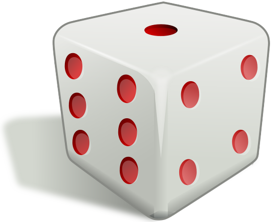
What is the probability that we roll a fair 6-sided die 4 times before finally rolling a 5 (on the 5^{th} roll)?

The probability of "success" (rolling a 5) is $\frac{1}{6}$.

$$\mathbb{P}(X = k) = (1 - p)^{k-1} p$$

$$\mathbb{P}(X = 5) = \left(1 - \frac{1}{6}\right)^{(5-1)} \left(\frac{1}{6}\right)$$

$$= \frac{625}{7776} \approx 0.0804$$



Uniform Distribution

With the continuous uniform distribution, all intervals of the same length on the distribution's "support" (between a and b) are equally probable.

- Probability Density Function (PDF): $f(x) = \frac{1}{b-a}$ for $x \in [a, b]$.
- Mean: $\frac{1}{2}(a + b)$.
- Variance: $\frac{1}{12}(b - a)^2$.

Examples of uniform distributions include:

- The probability that a clock battery dies at a specific time of day.
- The usual default for pseudo-random number generators.

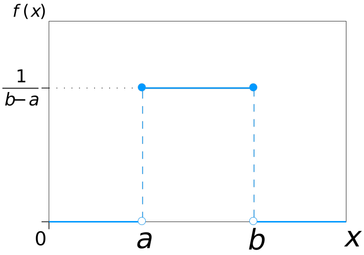


Image Source

Uniform Distribution Example (Clock)

What is the probability that, if you arrive at Big Ben at any given time and wait for 5 minutes, you'll hear the bells that mark the hour?

Uniform Distribution Example (Clock)

What is the probability that, if you arrive at Big Ben at any given time and wait for 5 minutes, you'll hear the bells that mark the hour?

The bells that mark the hour go off every 60 minutes, and if you wait 5 minutes, the probability of hearing the bells is $\frac{5}{60} = \frac{1}{12}$.

Normal / Gaussian Distribution

The normal / Gaussian distribution is the famous "bell curve" that appears in countless applications as a consequence of the central limit theorem.

- $f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- Mean: μ.
- Variance: σ^2.

Examples include:

- The distribution of women's heights.
- The distribution of standardized test scores.
- The distribution of soup can weights at a soup can factory.

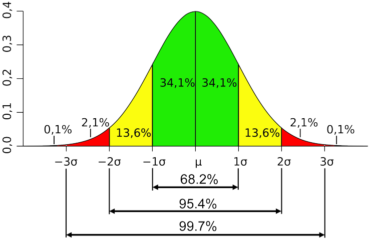


Image Source

Normal Distribution Example: Blood Pressure

The average systolic blood pressure for an adult is 125 mm Hg, with standard deviation 10 mm Hg. What percentage of people have "high blood pressure"?

Normal Distribution Example: Blood Pressure

We could solve this by integration:

$$\int_{x_0}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \int_{140}^{\infty} \frac{1}{10\sqrt{2\pi}} e^{\frac{-(x-125)^2}{2(10)^2}} dx$$

but it's much simpler to calculate a z-score:

$$z = \frac{x_0 - \mu}{\sigma} = \frac{140 - 125}{10} = 1.5$$

From a z-score table:

$$\mathbb{P}\{z > 1.5\} = 0.0668$$

So 6.68% of adults have high blood pressure.

Exponential Distribution

The exponential distribution describes the time between events in a *Poisson process*, which occur at a constant average rate.

- $f(x|\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$.
- Mean: $\frac{1}{\lambda}$.
- Variance: $\frac{1}{\lambda^2}$.

Examples include:

- The time between seeing two shooting stars.
- Distance between potholes on the street.
- The number of miles driven before a car battery wears out.

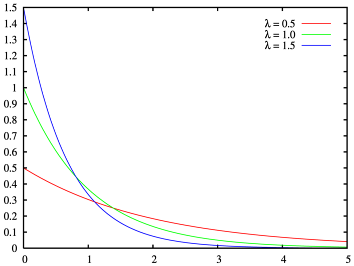


Image Source

Exponential Distribution Example (Light Bulb)

An incandescent light bulb lasts, on average, 1000 hours before burning out. What is the probability that a given bulb lasts over 2000 hours?

Exponential Distribution Example (Light Bulb)

For a light bulb, we know $\mu = \frac{1}{\lambda} = 1000$. From this, the probability density function f is

$$f(x) = \lambda e^{-\lambda x} = \frac{1}{1000} e^{\frac{-x}{1000}}$$

We can find the probability that the bulb lasts more than 2000 hours using integration:

$$\begin{aligned} \int_{2000}^{\infty} \frac{1}{1000} e^{\frac{-x}{1000}} dx &= \left[e^{\frac{-x}{1000}} \right] \Big|_{2000}^{\infty} \\ &= 0 - \left(-e^{\frac{-2000}{1000}} \right) \\ &= e^{-2} \approx 0.1353 \end{aligned}$$

There is an approximately 13.53% chance of the light bulb lasting more than 2000 hours.

χ^2 Distribution

The χ^2 (chi-square) distribution is the distribution of a sum of squares of k independent normal random variables. We often use this in hypothesis testing or the construction of confidence intervals.

- $f(x|k) = \frac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}$.
- Mean: k.
- Variance: $2k$.

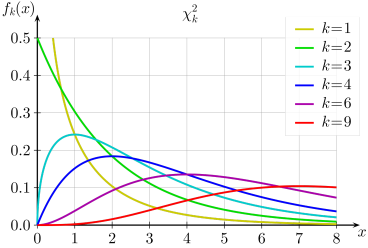


Image Source

Central Limit Theorem

The central limit theorem (CLT) states that the mean of a large number of independent random variables will have a distribution that is approximately normal / Gaussian. It is one of the most important results in probability and statistics, and explains why so many observations in the real world seem to follow a familiar "bell-curve" pattern.

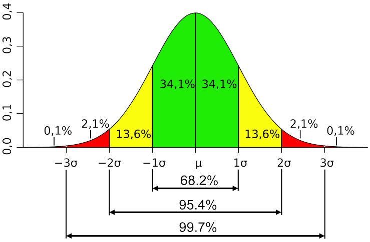


Image Source

Central Limit Theorem

More formally, if we have a collection if independent random variables $\{X_1, X_2, \ldots, X_n\}$, each with mean μ and standard deviation σ, then for large n:

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

where $\mathcal{N}(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ, and $\overline{X}_n$ is the mean of all n random variables $\{X_1, \ldots, X_n\}$.