Mathematical Foundation of computer science

Mathematics (Indiana Institute of Technology)

UNIT 1

Probability mass, density, and cumulative distribution functions, Parametric families of distributions, Expected value, variance, conditional expectation, Applications of the univariate and multivariate Central Limit Theorem, Probabilistic inequalities, Markovchains.

PROBABILITY MASS

A function which maps the members of a sample space to probabilities of their occurrence is known as probability mass function. The probability mass function is defined for discrete random variables, that is, random variable having discrete values. Suppose, a die is rolled. Rolling of die can have any one of six outcomes {1, 2, 3, 4, 5, 6} {1, 2, 3, 4, 5, 6}. The probability attached to each outcome can be defined using a function known as probability mass function. In this case, the probability mass function, p.m.fp.m.f = 1616.

Formal Definition

Like most statistical terms, there's the informal definition, and then there's the formal one:

The probability mass function, f(x) = P(X = x), of a discrete random variable X has the following properties:

1. All probabilities are positive: fx(x) ≥ 0.
2. Any event in the distribution (e.g. "scoring between 20 and 30") has a probability of happening of between 0 and 1 (e.g. 0% and 100%).
3. The sum of all probabilities is 100% (i.e. 1 as a decimal): Σfx(x) = 1.
4. AN individual probability is found by adding up the x-values in event A. P(X E A) =

$$\sum_{x \in A} f(x)$$

Example

Experiment :: Toss a fair coin 3 times

Sample Space :: S = {HHH, HHT, HT H, T HH, HT T, T HT, T T H, T T T }

Random variable X is the number of tosses.

Thus X : S → R looks like this

X(HHH) = 3

X(HHT) = X(HT H) = X(T HH) = 2

X(HT T) = X(T HT) = X(T T H) = 1

X(T T T) = 0

Thus, Range(X) = {0, 1, 2, 3} and

P(X = 0) = 1 8 , P(X = 1) = 3 8 , P(X = 2) = 3 8 , P(X = 3) = 1 8

Hence the probability mass function is given by

pX(0) = 1 8 , pX(1) = 3 8 , pX(2) = 3 8 , pX(3) = 1 8

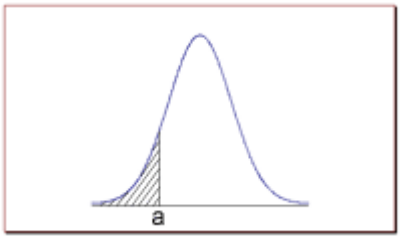
Probability Density Function

For continuous random variables, the function mapping the values of the variable in a certain interval to probability of its occurrence is known as probability density function.For continuous random variables, the probability density function is the function which to be integrated on the given interval to get the probability of getting a value within that interval. The probability density function for getting the value of a random variable between 33 and 1313 can be represented as $P(3 < x < 13)$P(3 < x < 13).

Most often, the equation used to describe a continuous probability distribution is called a **probability density function** . Sometimes, it is referred to as a **density function**, a **PDF**, or a **pdf**. For a continuous probability distribution, the density function has the following properties:

- Since the continuous random variable is defined over a continuous range of values (called the **domain** of the variable), the graph of the density function will also be continuous over that range.
- The area bounded by the curve of the density function and the x-axis is equal to 1, when computed over the domain of the variable.
- The probability that a random variable assumes a value between a and b is equal to the area under the density function bounded by a and b.

For example, consider the probability density function shown in the graph below. Suppose we wanted to know the probability that the random variable X was less than or equal to a. The probability that X is less than or equal to a is equal to the area under the curve bounded by a and minus infinity - as indicated by the shaded area.

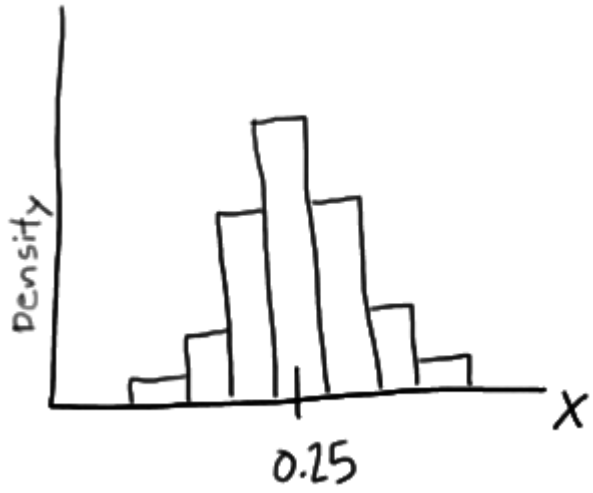


Note: The shaded area in the graph represents the probability that the random variable X is less than or equal to a. This is a cumulative probability . However, the probability that X is exactly equal to a would be zero. A continuous random variable can take on an infinite number of values. The probability that it will equal a specific value (such as a) is always zero.

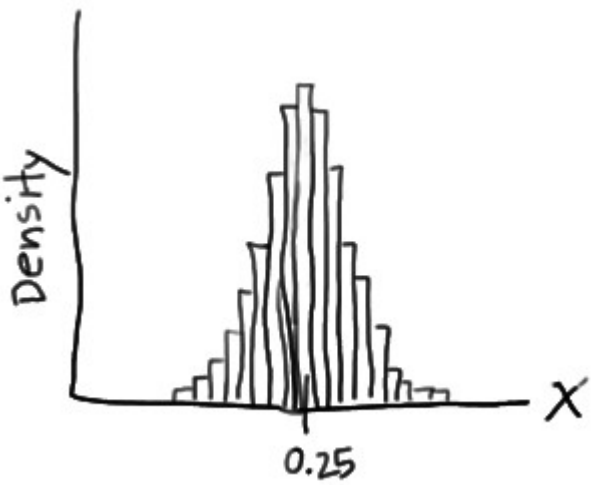
EXAMPLE

Even though a fast-food chain might advertise a hamburger as weighing a quarter-pound, you can well imagine that it is not exactly 0.25 pounds. One randomly selected hamburger might weigh 0.23 pounds while another might weigh 0.27 pounds. What is the probability that a randomly selected hamburger weighs between 0.20 and 0.30 pounds? That is, if we let X denote the weight of a randomly selected quarter-pound hamburger in pounds, what is $P(0.20 < X < 0.30)$?

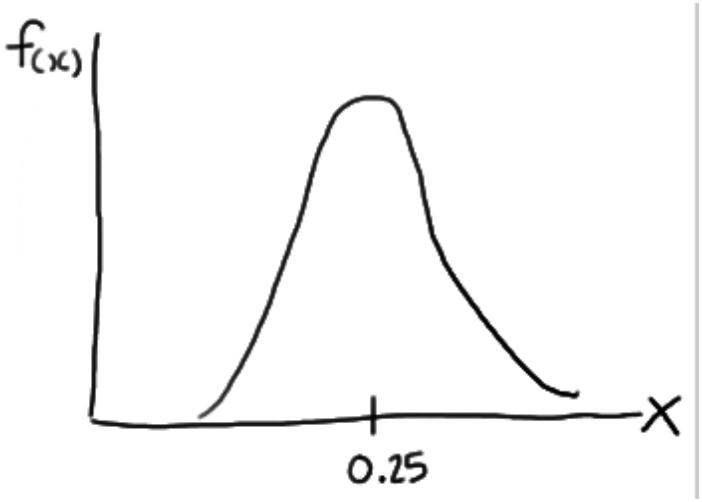
Solution. In reality, I'm not particularly interested in using this example just so that you'll know whether or not you've been ripped off the next time you order a hamburger! Instead, I'm interested in using the example to illustrate the idea behind a probability density function. Now, you could imagine randomly selecting, let's say, 100 hamburgers advertised to weigh a quarter-pound. If you weighed the 100 hamburgers, and created a density histogram of the resulting weights, perhaps the histogram might look something like this:



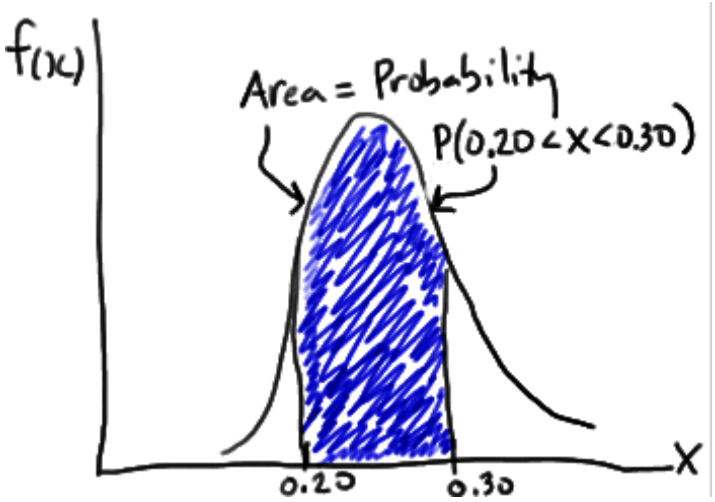
In this case, the histogram illustrates that most of the sampled hamburgers do indeed weigh close to 0.25 pounds, but some are a bit more and some a bit less. Now, what if we decreased the length of the class interval on that density histogram? Then, the density histogram would look something like this:



Now, what if we pushed this further and decreased the intervals even more? You can imagine that the intervals would eventually get so small that we could represent the probability distribution of *X,* not as a density histogram, but rather as a curve (by connecting the "dots" at the tops of the tiny tiny tiny rectangles) that, in this case, might look like this:



Such a curve is denoted ***f*(*x*)** and is called a (continuous) **probability density function**.
Now, you might recall that a density histogram is defined so that the area of each rectangle equals the relative frequency of the corresponding class, and the area of the entire histogram equals 1. That suggests then that finding the probability that a continuous random variable *X* falls in some interval of values involves finding the area under the curve *f*(*x*) sandwiched by the endpoints of the interval. In the case of this example, the probability that a randomly selected hamburger weighs between 0.20 and 0.30 pounds is then this area:



Now that we've motivated the idea behind a probability density function for a continuous random variable, let's now go and formally define it.

Cumulative Distribution Function

The PMF is one way to describe the distribution of a discrete random variable. As we will see later on, PMF cannot be defined for continuous random variables. The cumulative distribution function (CDF) of a random variable is another method to describe the distribution of random variables. The advantage of the CDF is that it can be defined for any kind of random variable (discrete, continuous, and mixed).

Definition

The cumulative distribution function (CDF) of random variable XX is defined as FX(x)=P(X≤x), for all x∈R

Suppose p(x) is a density function for a quantity.

The cumulative distribution function (cdf) for the quantity is defined as

Gives:

$$P(x) = \int_{-\infty}^{x} p(t)\,dt$$

• The proportion of population with value less than x

• The probability of having a value less than x.

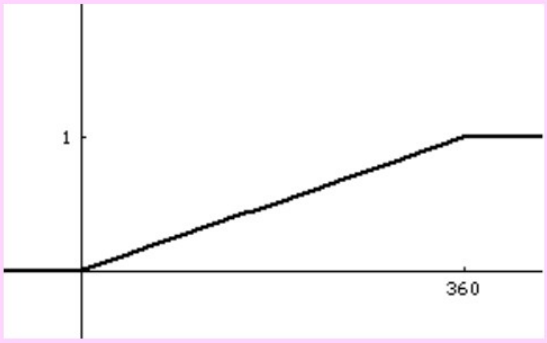
Example: A Spinner Last class: A spinner that could take on any value 0º ≤ x ≤ 360º. Density function: p(x) = 1/360 if 0 ≤ x ≤ 360, and 0 everywhere else.

Last class: A spinner that could take on any value 0º ≤ x ≤ 360º. Density function: p(x) = 1/360 if 0 ≤ x ≤ 360, and 0 everywhere else. CDF:

$$P(x) = \int_{-\infty}^{x} p(t)\,dt = \begin{cases} 0, \text{ if } x < 0 \\ \dfrac{x}{360}, \text{ if } 0 \le x \le 360 \\ 1, \text{ if } x > 360 \end{cases}$$

Density Function:

Cumulative Distribution Function:



Properties of CDFs

• P(x) is the probability of values less than x

– If P(x) is the cdf for the age in months of fish in a lake, then P(10) is the probability a random fish is 10 months or younger.

• P(x) goes to 0 as x gets smaller:

$$\lim_{x \to -\infty} P(x) = 0$$

(In many cases, we may reach 0.)

• Conversely,

$$\lim_{x \to \infty} P(x) = 1$$

• P(x) is non-decreasing.

– The derivative is a density function, which cannot be negative.

– Also, P(4) can't be less than P(3), for example.

Example Someone claims this is the CDF for grades on the 2015 final exam.

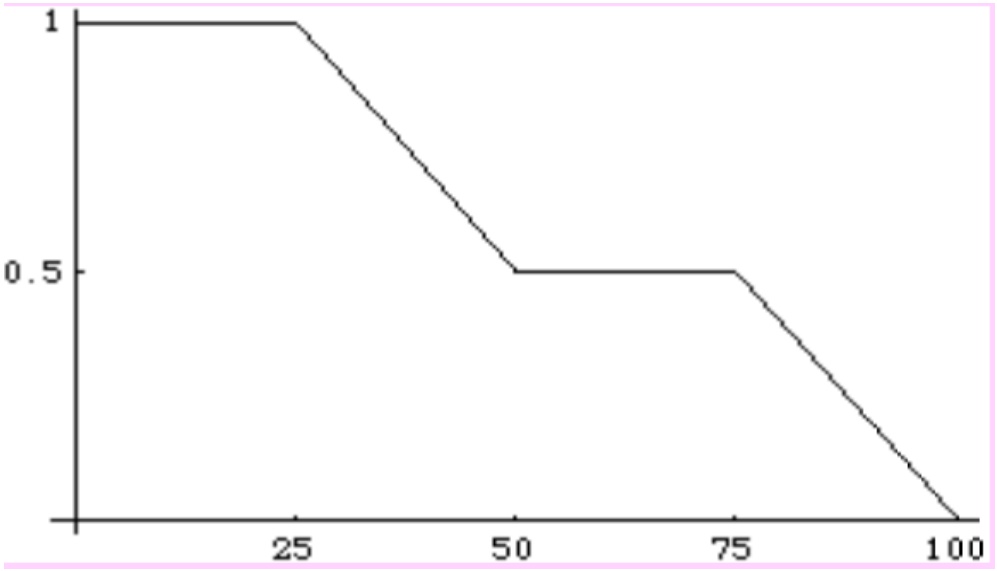
Probability a random student scored….

• 25 or lower?

1, or 100%.

• 50 or lower?

0.5, or 50%.



Conclusion?

They're lying! This cannot be a cumulative distribution function! (It decreases.)

Point Estimation for Parametric Families of Probability Distributions

Parametric Families

We now shift gears to discuss the statistical idea of point estimation. We will be consider the problem of parametric point estimation, so we will first need to understand what is a parametric family of probability distributions.

Here a parameter space Θ will be a subset of R k for some k.

Definition 2. A parametric family of probability distributions is a collection of probability density functions (or probability mass functions) on R n indexed by the parameter space Θ, that is, a collection of densities of the form {f(x; θ) : θ ∈ Θ}.

Given a parametric family, each θ ∈ Θ uniquely specifies a probability density function f(x; θ).

Example 1 (Normal family). The family of normal probability densities has parameter space Θ = R × (0,∞). In this case, the parameter is the ordered pair θ = (μ, σ2), and the density specified by θ is (in the case of an i.i.d. sample (X1, . . . , Xn) of size n)

$$f(x;\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Suppose that the distribution of the random vector X has a density belonging to a parametric family, that is, X ∼ f(x; θ) for some θ ∈ Θ. Given a function g : R n → R, we write Eθ(g(X)) to

indicate that we are taking an expectation with respect to the density f(x; θ). Similarly, we write Pθ(X ∈ A) to indicate we are computing a probability using the density f(x; θ). To be precise,

$$P_\theta(\boldsymbol{X} \in A) = \int_A f(\boldsymbol{x};\theta)d\boldsymbol{x}$$

$$E_\theta(g(\boldsymbol{X})) = \int_{\mathbb{R}^n} g(\boldsymbol{x})f(\boldsymbol{x};\theta)\boldsymbol{x}\,.$$

A parametric family can have more than one parameterization. For example, we can parameterize the exponential family by

$$\mu \mapsto \frac{1}{\mu}e^{-x/\mu}\mathbf{1}\left\{x \geq 0\right\}\,, \quad \mu > 0\,.$$

Alternatively, it is sometimes parameterized by

$$\lambda \mapsto \lambda e^{-\lambda x}\mathbf{1}\left\{x \geq 0\right\}\,, \quad \lambda > 0\,.$$

When we talk about a parametric family of probability distributions, we should be sure to specify explicitly which parameterization we are using.

EXPECTED VALUE

What is Expected Value?

Expected value is exactly what you might think it means intuitively: the **return you can expect for some kind of action**, like how many questions you might get right if you guess on a multiple choice test.

For example, if you take a 20 question multiple-choice test with A,B,C,D as the answers, and you guess all “A”, then you can expect to get 25% right (5 out of 20). The **math** behind this kind of expected value is:

The probability (P) of getting a question right if you guess: .25

The number of questions on the test (n)*: 20
P x n = .25 x 20 = 5
**You might see this as X instead.*
This type of expected value is called an expected value for a binomial random variable. It's binomial because there are only two possible outcomes: you get the answer right, or you get the answer wrong.

The mean of the discrete random variable X is also called the **expected value** of X. Notationally, the expected value of X is denoted by E(X). Use the following formula to compute the mean of a discrete random variable.

$$E(X) = \Sigma [x_i * P(x_i)]$$

where x_i is the value of the random variable for outcome i, and $P(x_i)$ is the probability that the random variable will be equal to outcome i.

VARIANCE

What is 'Variance'

Variance is a measurement of the spread between numbers in a data set. The variance measures how far each number in the set is from the mean. Variance is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set.

BREAKING DOWN 'Variance'

Variance is used in statistics for probability distribution. Since variance measures the variability (volatility) from an average or mean and volatility is a measure of risk, the variance statistic can help determine the risk an investor might assume when purchasing a specific security. A variance value of zero indicates that all values within a set of numbers are identical; all variances that are non-zero will be positive numbers. A large variance indicates that numbers in the set are far from the mean and each other, while a small variance indicates the opposite

Advantages and Disadvantages of Using Variance

Statisticians use variance to see how individual numbers relate to each other within a data set, rather than using broader mathematical techniques such as arranging numbers into quartiles. A drawback to variance is that it gives added weight to numbers far from the mean (outliers), since squaring these numbers can skew interpretations of the data. The advantage of variance is that it treats all deviations from the mean the same regardless of direction; as a result, the squared deviations cannot sum to zero and give the appearance of

no variability at all in the data. The drawback of variance is that it is not easily interpreted, and the square root of its value is usually taken to get the standard deviation of the data set in question.

Variance - Example

A study has 100 people perform a simple speed task during 80 trials. For each participant, 80 reaction times (in seconds) are thus recorded. Part of these data are shown below.

	name	time1	time2	time3	tim
1	Jason	2.349	1.602	1.441	
2	Taylor	2.427	2.031	2.074	
3	Evelyn	1.701	1.688	1.879	
4	Chase	1.850	1.523	1.380	

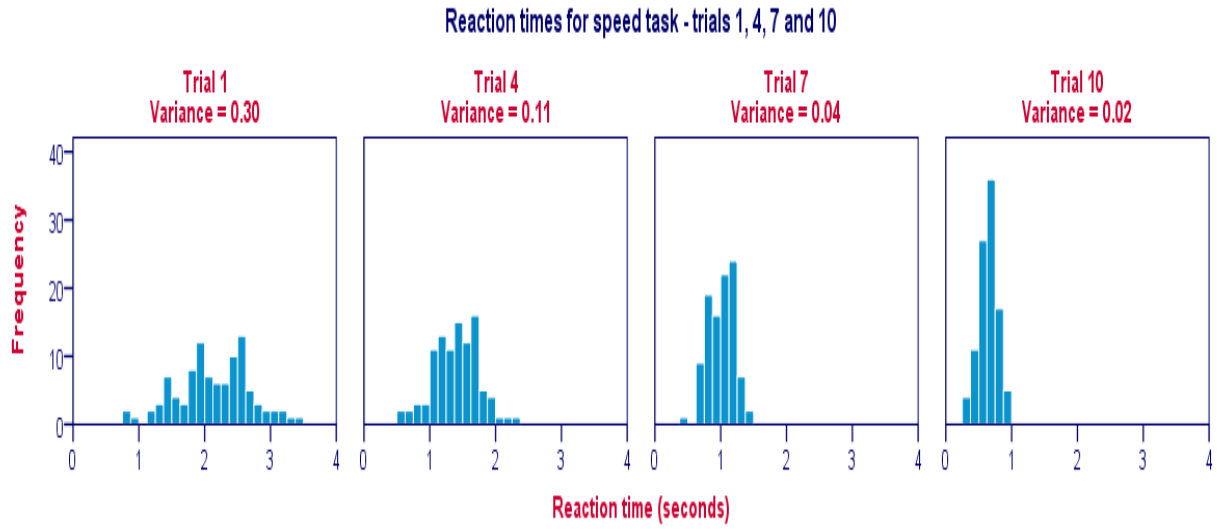
In studies like these, we typically see that people get faster as they perform the speed task more often. That is, the average reaction time tends to decrease over trials.
Also, reaction times will typically *vary* less between different people insofar as they perform the task more often. Technically, we say that the **variance decreases over trials**. The table below illustrates this for trials 1,4,7 and 10.

Reaction times for speed task - trials 1, 4, 7 and 10

Trial	N	Mean	Std. Deviation	Variance
1	100	2.14	.55	.30
4	100	1.40	.33	.11
7	100	1.01	.19	.04
10	100	.65	.14	.02

Variance and Histogram

A great way to visualize the data from our previous table is a histogram for each trial. Like so, the figure below illustrates that participants got faster over trials; from trial 1 to trial 10 the histogram bars move leftwards, towards 0 seconds.



A second finding is that the histograms become narrower (and therefore higher) as we move from trial 1 to trial 10; this illustrates that reaction times vary less and less between our participants as the experiment progresses. The variance decreases over trials.

Variance - Population Formula

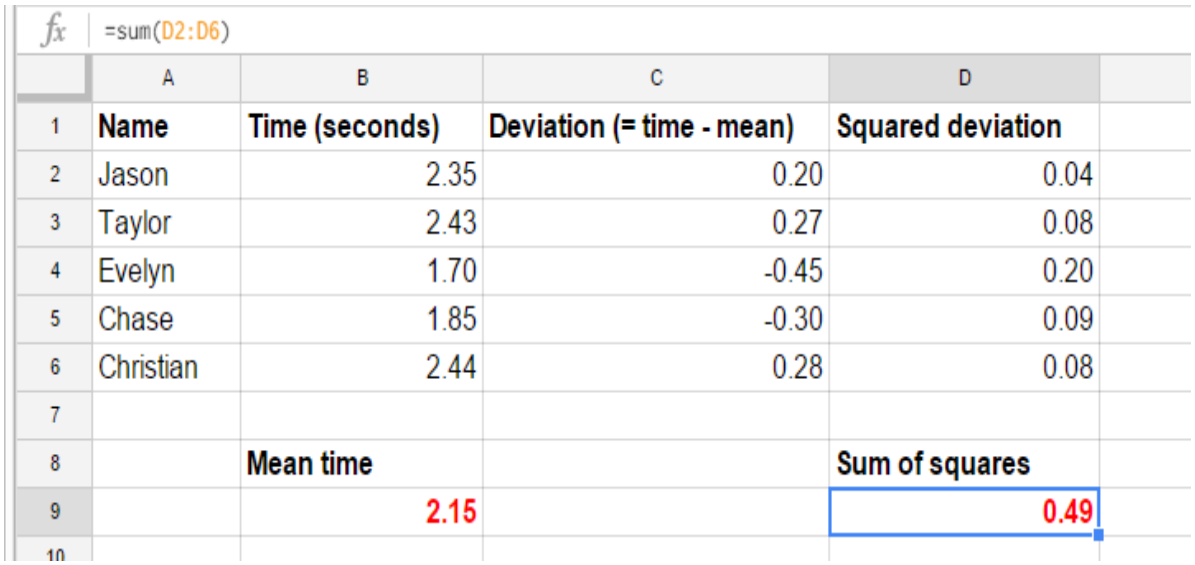
A basic formula for calculating the variance is

$$S^2=\sum(X-\bar{X})^2/n$$

We recommend you try to understand what this formula does because this helps a lot in understanding ANOVA (= analysis of variance). We'll therefore demonstrate it on a mere handful of data.

Variance - GoogleSheets

For the sake of simplicity, we'll cut down our data to the first trial for the first 5 participants. These 5 reaction times -and a manual calculation of their variance- are in this GoogleSheet.



=sum(D2:D6)

	A	B	C	D
1	**Name**	**Time (seconds)**	**Deviation (= time - mean)**	**Squared deviation**
2	Jason	2.35	0.20	0.04
3	Taylor	2.43	0.27	0.08
4	Evelyn	1.70	-0.45	0.20
5	Chase	1.85	-0.30	0.09
6	Christian	2.44	0.28	0.08
7				
8		**Mean time**		**Sum of squares**
9		**2.15**		**0.49**
10				

Variance - Calculation Steps

The formulas in the GoogleSheet show precisely how to calculate a variance. The basic steps are

1. calculate the **mean** reaction time (2.15);
2. calculate **deviation scores** (reaction time minus mean reaction time);
3. calculate **squared deviation** scores;
4. add squared deviation scores. The result (0.49) is a **sum of squares**, the main building block of ANOVA;
5. **divide** the sum of squares by the number of observations (5 reaction times).

Alternatively, calculate a variance by typing =VARP(B2:B6) in some cell (B2:B6 are the cells that hold our 5 reaction times). VARP is short for “variance population”. OpenOffice and MS Excel contain similar formulas.

Variance - Sample Formula

Similarly to the standard deviation, if our data are a simple random sample from a much larger population, the aforementioned formula will systematically underestimate the population variance. In this case we'll use a slightly different formula:

$$S_2=\sum(X-X^{\overline{\quad}})_2 n-1 S2=\sum(X-X^{-})2n-1$$

Which formula to use thus depends on our data: do they contain the entire population we'd like to investigate or are they a mere sample from this population? Since our 100 participants are clearly a sample, we'll use the sample formula. In GoogleSheets, typing =VAR(B2:B6) in some cell will return the sample variance.

Variance in SPSS

Insofar as we know, the formula for the population variance is completely absent from SPSS and we consider this a serious flaw. Instead, **SPSS always uses the sample formula**.* Relevant output is shown below.

Descriptive Statistics

	N	Mean	Std. Deviation	Variance
Time (seconds) for trial 1	5	2.15	.35	.12

Regarding this output table, also note that the variance is indeed the squared standard deviation (apart from rounding).

Regarding the variance, that's about it. We hope you found this tutorial helpful in understanding what a variance is.

Conditional expectation

The conditional expectation (or conditional mean, or conditional expected value) of a random variable is the expected value of the random variable itself, computed with respect to its conditional probability distribution.

As in the case of the expected value, giving a completely rigorous definition of conditional expected value requires a complicated mathematical apparatus. To make things simpler, we do not give a completely rigorous definition in this lecture. We rather give an informal definition and we show how conditional expectation can be computed. In particular, we discuss how to compute the expected value of a random variable x when we observe the realization of another random variable y, i.e. when we receive the information that $Y= y$. The expected value of x conditional on the information that $Y= y$ is called conditional expectation of x given $Y= y$.

Definition

The following informal definition is very similar to the definition of expected value we have given in the lecture entitled Expected value.

Definition (informal) Let x and y be two random variables.
The **conditional expectation** of x given Y= y is the weighted average of the values that x can take on, where each possible value is weighted by its respective conditional probability (conditional on the information that Y= y).

The expectation of a random variable x conditional on Y= y is denoted by E[X[Y = y].

Conditional expectation of a discrete random variable

In the case in which x and y are two discrete random variables and, considered together, they form adiscrete random vector, the formula for computing the conditional expectation of x given Y= y is a straightforward implementation of the above informal definition of conditional expectation: the weights of the average are given by the conditional probability mass function of x .

Definition Let x and y be two discrete random variables. Let R_X be the support of x and let $p_{X|Y=y}(x)$ be the conditional probability mass function of x given Y= y. The conditional expectation of x given Y= y is provided that

$$\sum_{x \in R_X} |x| p_{X|Y=y}(x) < \infty$$

If you do not understand the symbol $\sum_{x \in R_X}$ and the finiteness condition above (absolute summability) go back to the lecture entitled Expected value, where they are explained.

Example Let the support of the random vector [X Y] be

$$R_{XY} = \left\{ \begin{bmatrix} 1 & 3 \end{bmatrix}, \begin{bmatrix} 2 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \end{bmatrix} \right\}$$

and its joint probability mass function be

$$p_{XY}(x,y) = \begin{cases} \frac{1}{3} & \text{if } x = 1 \text{ and } y = 3 \\ \frac{1}{3} & \text{if } x = 2 \text{ and } y = 0 \\ \frac{1}{3} & \text{if } x = 0 \text{ and } y = 0 \\ 0 & \text{otherwise} \end{cases}$$

Let us compute the conditional probability mass function of x given y =0 . The marginal probability mass function of y evaluated at y=0 is The support of x is

$$R_X = \{0, 1, 2\}$$

Thus, the conditional probability mass function of X given Y=0 is The conditional expectation of X given Y is

$$\begin{aligned} \mathrm{E}[X|Y=0] &= 0 \cdot p_{X|Y=0}(0) + 1 \cdot p_{X|Y=0}(1) + 2 \cdot p_{X|Y=0}(2) \\ &= 0 \cdot \frac{1}{2} + 1 \cdot 0 + 2 \cdot \frac{1}{2} = 1 \end{aligned}$$

Conditional expectation of an absolutely continuous random variable

When x and y are absolutely continuous random variables, forming an absolutely continuous random vector, the formula for computing the conditional expectation of x givenY= y involves an integral, which can be thought of as the limiting case of the summation

$\sum_{x \in R_X} x p_{X|Y=y}(x)$ found in the discrete case above.

Definition Let x and y be two absolutely continuous random variables. Let R_X be the support of x and let $f_{X|Y=y}(x)$ be the conditional

probability density function of given Y= y The conditional expectation of given Y= y
is

$$\mathrm{E}[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y=y}(x)dx$$

provided that

$$\int_{-\infty}^{\infty} |x| f_{X|Y=y}(x)dx < \infty$$

If you do not understand why an integration is required and why the finiteness condition above (absolute integrability) is imposed, you can find an explanation in the lecture entitled Expected value.

Example Let the support of the random vector [X Y] be

$$R_{XY} = [0,\infty) \times [2,4]$$

and its joint probability density function be

$$f_{XY}(x,y) = \begin{cases} \frac{1}{2} y \exp(-xy) & \text{if } x \in [0,\infty) \text{ and } y \in [2,4] \\ 0 & \text{otherwise} \end{cases}$$

Let us compute the conditional probability density function of x given Y=2 . The support of Y is

$$R_Y = [2,4]$$

When $y \notin [2,4]$, the marginal probability density function of Y is 0 ; when $y \in [2,4]$, the marginal probability density function is

$$
\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dx \\
&= \int_0^{\infty} \frac{1}{2} y \exp(-xy)dx \\
&= \frac{1}{2}[-\exp(-xy)]_0^{\infty} \\
&= \frac{1}{2}[0-(-1)] \\
&= \frac{1}{2}
\end{aligned}
$$

Thus, the marginal probability density function of Y is

$$
f_Y(y) = \begin{cases} \frac{1}{2} & \text{if } y \in [2,4] \\ 0 & \text{otherwise} \end{cases}
$$

When evaluated at ,y=2 it is

$$
f_Y(2) = \frac{1}{2}
$$

The support of is

$$
R_X = [0,\infty)
$$

Thus, the conditional probability density function of x given Y=2 is

$$
f_{X|Y=2}(x) = \frac{f_{XY}(x,2)}{f_Y(2)} = \begin{cases} 2\exp(-2x) & \text{if } x \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}
$$

The conditional expectation of given Y= y is

$$
\begin{aligned}
\mathrm{E}[X|Y=2] &= \int_{-\infty}^{\infty} x f_{X|Y=2}(x)\,dx \\
&= \int_{0}^{\infty} x \cdot 2\exp(-2x)\,dx \\
&= \frac{1}{2}\int_{0}^{\infty} t\exp(-t)\,dt \qquad \textit{(by changing variables: } t = 2x\textit{)} \\
&= \frac{1}{2}\left\{[-t\exp(-t)]_0^{\infty} + \int_{0}^{\infty}\exp(-t)\,dt\right\} \qquad \textit{(integrating by parts)} \\
&= \frac{1}{2}\{0 - 0 + [-\exp(-t)]_0^{\infty}\} \\
&= \frac{1}{2}\{0 + 1\} = \frac{1}{2}
\end{aligned}
$$

Conditional expectation in general

The general formula for computing the conditional expectation of given does not require that the two variables form a discrete or an absolutely continuous random vector, but it is applicable to any random vector.

Definition Let $F_{X|Y=y}(x)$ be the conditional distribution function of x given Y= y. The conditional expectation of x given Y= y is

$$\mathrm{E}[X|Y=y] = \int_{-\infty}^{\infty} x\,dF_{X|Y=y}(x)$$

where the integral is a Riemann-Stieltjes integral and the expected value exists and is well-defined only as long as the integral is well-defined.

The above formula follows the same logic of the formula for the expected value

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x\,dF_X(x)$$

with the only difference that the unconditional distribution function $F_X(x)$ has now been replaced with the conditional distribution function $F_{X|Y=y}(x)$. . The reader who feels unfamiliar with this formula

can go back to the lecture entitled Expected value and read an intuitive introduction to the Riemann-Stieltjes integral and its use in probability theory.

More details

The following subsections contain more details about conditional expectation.

Properties of conditional expectation

From the above sections, it should be clear that the conditional expectation is computed exactly as the expected value, with the only difference that probabilities and probability densities are replaced by conditional probabilities and conditional probability densities. Therefore, the properties enjoyed by the expected value, such as linearity, are also enjoyed by the conditional expectation. For an exposition of the properties of the expected value, you can go to the lecture entitled Properties of the expected value.

Law of iterated expectations

Quite obviously, before knowing the realization of , the conditional expectation of given is unknown and can itself be regarded as a random variable. We denote it by E[X[Y] In other words, is a random variable such that its realization equals E[X]Y=Y] when is the realization of Y.

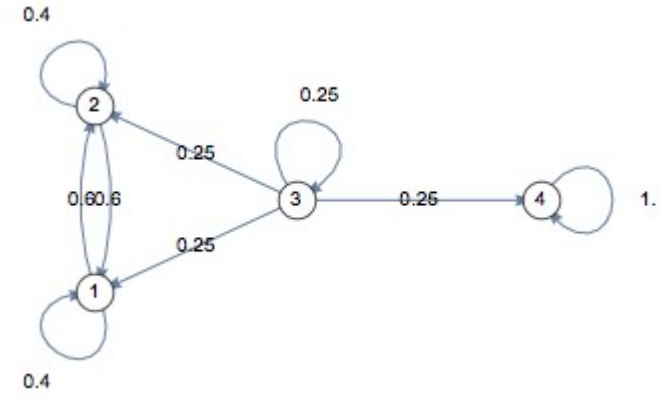
This random variable satisfies a very important property, known as **law of iterated expectations**: E[E[X[Y]]=[EX]

Proof

A **Markov chain** is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules. The defining characteristic of a Markov chain is that no matter *how* the process arrived at its present state, the possible future states are fixed. In other words, the probability of transitioning to any particular state is dependent solely on the current state and time elapsed. The **state space**, or set of all possible states, can be anything: letters, numbers, weather conditions, baseball scores, or stock performances.

Markov chains may be modeled by finite state machines, and random walks provide a prolific example of their usefulness in mathematics. They arise broadly in statistical and information-theoretical contexts and are widely employed in economics, game theory, queueing (communication) theory, genetics, and finance. While it is possible to discuss Markov chains with any size of state space, the initial theory and most applications are focused on cases with a finite (or countably infinite) number of states.

Many uses of Markov chains require proficiency with common matrix methods.

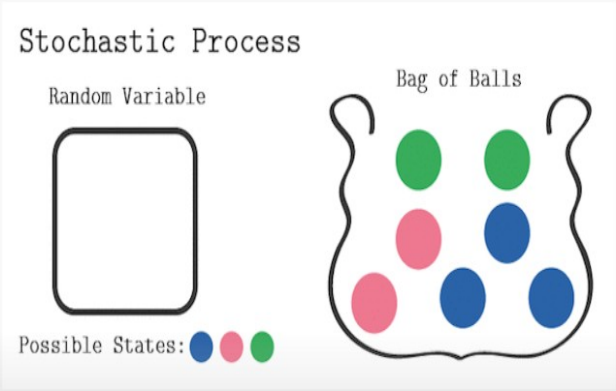


Basic Concept

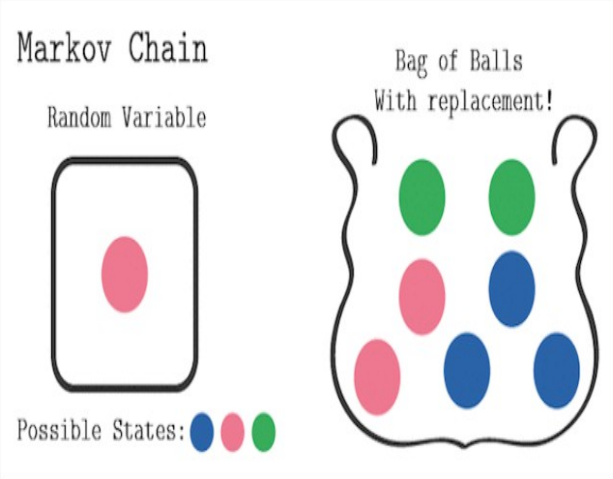
A Markov chain is a stochastic process, but it differs from a general stochastic process in that a Markov chain must be "memory-less". That is, (the probability of) future actions are not dependent upon the steps that led up to the present state. This is called the **Markov property**. While the theory of Markov chains is important precisely because so many "everyday" processes satisfy the Markov property, there are many common examples of stochastic properties that do not satisfy the Markov property.

EXAMPLE

A common probability question asks what is the probability of getting a certain color ball, when selecting uniformly and at random from a bag of multicolored balls. It could also ask what the probability of the next ball is, and so on. In such a way, a stochastic process begins to exist with color for the random variable, and it does not satisfy the Markov property. Depending upon which balls are removed, the probability of getting a certain color ball later may be drastically different.



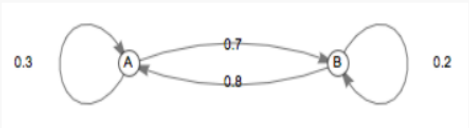
A variant of the same question asks once again for ball color, but it allows replacement each time a ball is drawn. Once again, this creates a stochastic process with color for the random variable. This process, however, **does** satisfy the Markov property. Can you figure out why?



In probability theory, the most immediate example is that of a *time-homogeneous Markov chain*, in which the probability of any state transition is independent of time. Such a process may be visualized with a labeled directed graph, for which the sum of the labels of any vertex's outgoing edges is 1.

EXAMPLE

A (time-homogeneous) Markov chain built on states A and B is depicted in the diagram below. What is the probability that a process beginning on A will be on B after 2 moves?



In order to move from A to B, the process must either stay on A the first move, then move to B the second move; or move to B the first move, then stay on B the second move. According to the diagram, the probability of that is $0.3 \cdot 0.7 + 0.7 \cdot 0.2 = \boxed{0.35.}$

Alternatively, the probability that the process will be on A after 2 moves is $0.3 \cdot 0.3 + 0.7 \cdot 0.8 = 0.65$. Since there are only two states in the chain, the process must be on B if it is not on A, and therefore, the probability that the process will be on B after 2 moves is $1 - 0.65 = \boxed{0.35.}$

In the language of conditional probability and random variables, a Markov chain is a sequence $X_0, X_1, X_2, \ldots$ of random variables satisfying the rule of conditional independence:

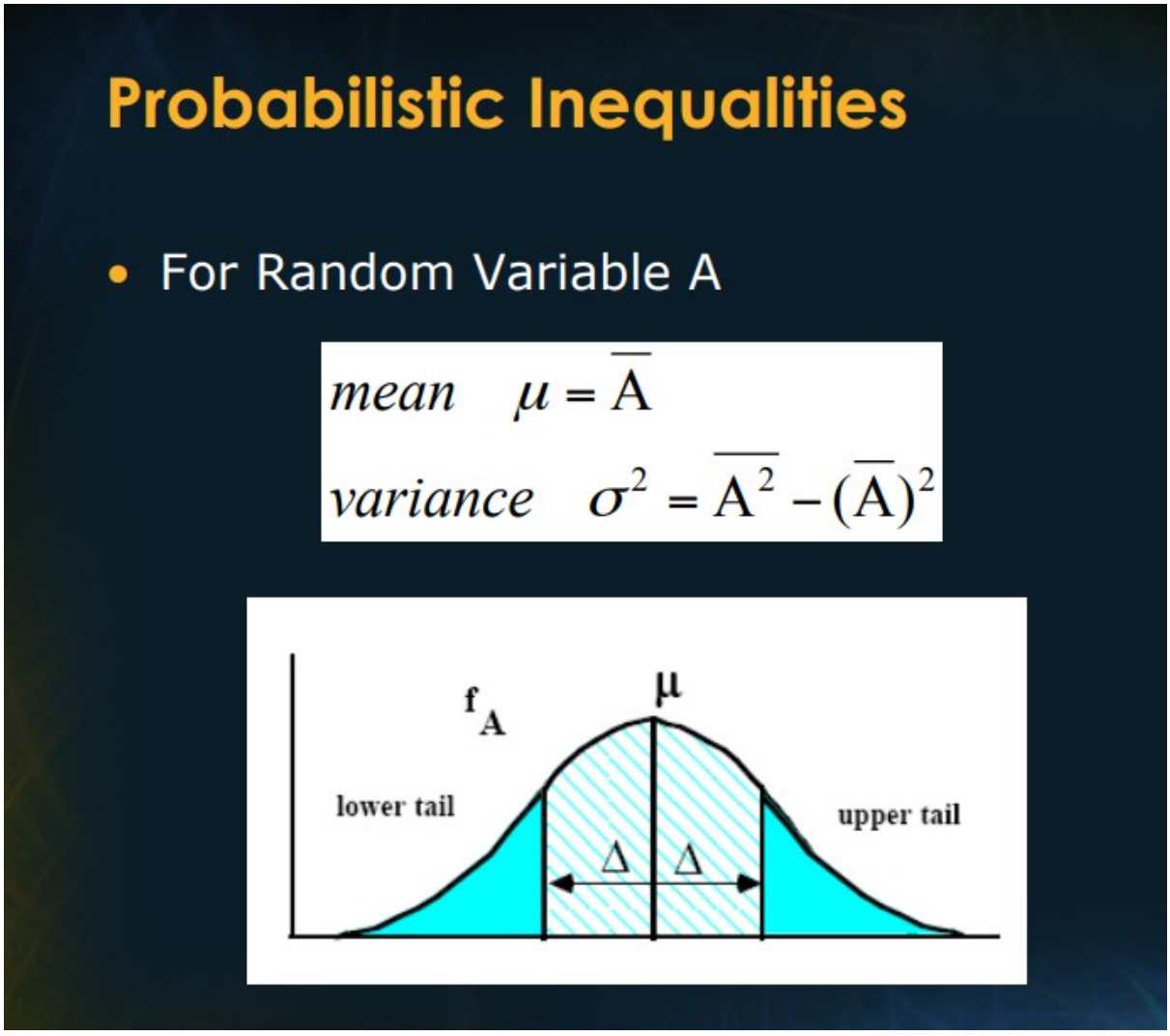
DEFINITION

The Markov Property.

For any positive integer n and possible states $i_0, i_1, \ldots, i_n$ of the random variables,

$$P(X_n = i_n \mid X_{n-1} = i_{n-1}) = P(X_n = i_n \mid X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}).$$

In other words, knowledge of the previous state is all that is necessary to determine the probability distribution of the current state. This definition is broader than the one explored above, as it allows for *non-stationary transition probabilities* and therefore *time-inhomogeneous Markov chains*; that is, as time goes on (steps increase), the probability of moving from one state to another may change.



Markov and Chebychev Probabilistic Inequalities

- Markov Inequality (uses only mean)

$$\text{Prob}\,(A \geq x) \leq \frac{\mu}{x}$$

- Chebychev Inequality

(uses mean and variance)

$$\text{Prob}\,(|A - \mu| \geq \Delta) \leq \frac{\sigma^2}{\Delta^2}$$