PROBABILITY THEORY

Math Tools Lab #7
November 1st, 2019

OUTLINE

(1) Summary statistics: review intuition for norms
(2) Code break 1
(3) Random variables and probability distributions: discrete and continuous
(4) Expected value, variance, and standard deviation
(5) Moments
(6) Code break 2
(7) Covariance and correlation
(8) **Code break 3**

SUMMARY STATISTICS

We've already looked at a few: mean, median, mode, etc.

What is the point of these? We want **a single number** that can represent the data well, which we can measure by looking at **the discrepancies from all data points using a metric**.

What metric should we use? Norms: $L_0, L_1, L_2, \ldots, L_p$

SUMMARY STATISTICS

We've already looked at a few: mean, median, mode, etc.

What is the point of these? We want **a single number** that can represent the data well, which we can measure by looking at **the discrepancies from all data points using a metric**.

What metric should we use? Norms: $L_0, L_1, L_2, \ldots, L_p$

Generalized L_p norm	$argmin_c[\frac{1}{N}\sum_{n=1}^{N} \vert x_n - c \vert^p]^{1/p}$
L_0 norm	mode
L_1 norm	median
L_2 norm	mean

SUMMARY STATISTICS

More intuitively, norm refers to **a function which assigns size to each vector in some vector space.** There are different ways to do this:

SUMMARY STATISTICS

More intuitively, norm refers to **a function which assigns size to each vector in some vector space**. There are different ways to do this:

L_0 Norm

Number of nonzero elements in a vector

$$X = [0,1]$$

$$||X||_0 = 1$$

SUMMARY STATISTICS

More intuitively, norm refers to **a function which assigns size to each vector in some vector space.** There are different ways to do this:

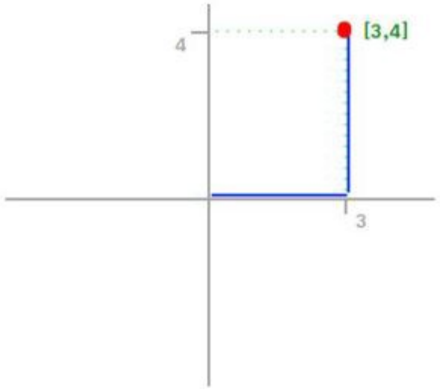
L_0 Norm

Number of nonzero elements in a vector

$X = [0,1]$

$||X||_0 = 1$

L_1 Norm



Manhattan distance or taxicab norm: sum of magnitude of vectors

$||X||_1 = |3| + |4| = 7$

SUMMARY STATISTICS

More intuitively, norm refers to **a function which assigns size to each vector in some vector space**. There are different ways to do this:

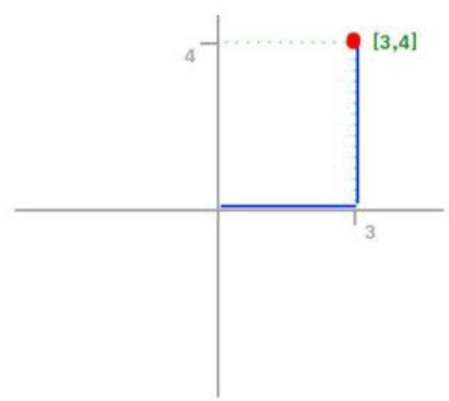
L_0 Norm

Number of nonzero elements in a vector

$$X = [0,1]$$

$$||X||_0 = 1$$

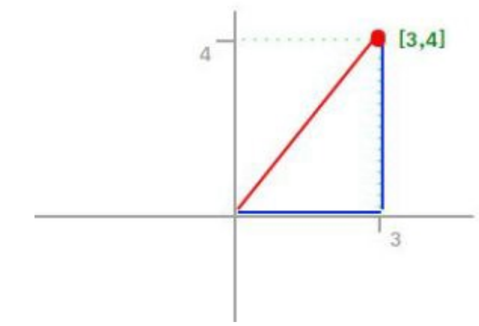
L_1 Norm



Manhattan distance or taxicab norm: sum of magnitude of vectors

$$||X||_1 = |3| + |4| = 7$$

L_2 Norm



Euclidean norm: shortest distance between points

$$||X||_2 = \sqrt{|3|^2 + |4|^2} = 25$$

SUMMARY STATISTICS

More intuitively, norm refers to **a function which assigns size to each vector in some vector space.** There are different ways to do this:

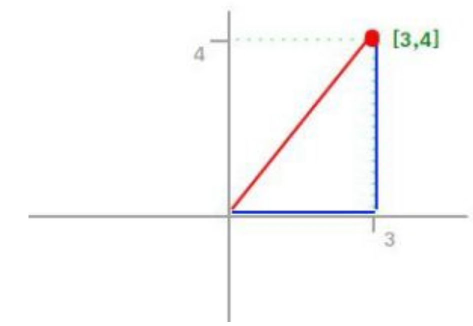
L_0 Norm

Number of nonzero elements in a vector

$$X = [0,1]$$

$$||X||_0 = 1$$

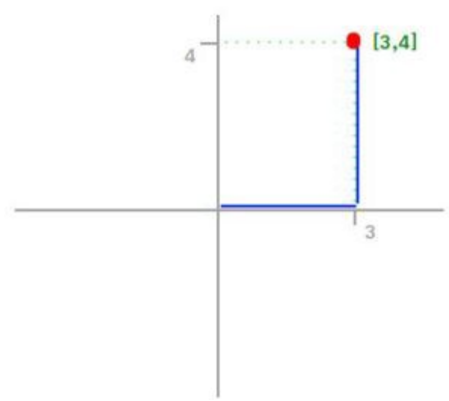
L_2 Norm



Euclidean norm: shortest distance between points

$$||X||_2 = \sqrt{|3|^2 + |4|^2} = 25$$

L_1 Norm



Manhattan distance or taxicab norm: sum of magnitude of vectors

$$||X||_1 = |3| + |4| = 7$$

L_∞ Norm

Largest magnitude element in a vector

$$X = [-6,4,2]$$

$$||X||_\infty = 6$$

CODE BREAK 1

DISCRETE PROBABILITY DISTRIBUTIONS: PMF

A discrete random variable is one that can equal a **finite number of values**. We can describe the distribution of a discrete r.v. using a **probability mass function (pmf)**, which must be non-negative for all inputs and sum to 1. A pmf describes the probability that a discrete r.v. X takes on a particular value, $P(X = x)$.

DISCRETE PROBABILITY DISTRIBUTIONS: PMF

A discrete random variable is one that can equal a **finite number of values**. We can describe the distribution of a discrete r.v. using a **probability mass function (pmf)**, which must be non-negative for all inputs and sum to 1. A pmf describes the probability that a discrete r.v. X takes on a particular value, $P(X = x)$.

Examples of discrete r.v.'s

(1) X = sum of two rolled dice
(2) X = maximum of two rolled dice
(3) X = # of coin flips until the 1st heads
(4) X = number of kids in a family

DISCRETE PROBABILITY DISTRIBUTIONS: PMF

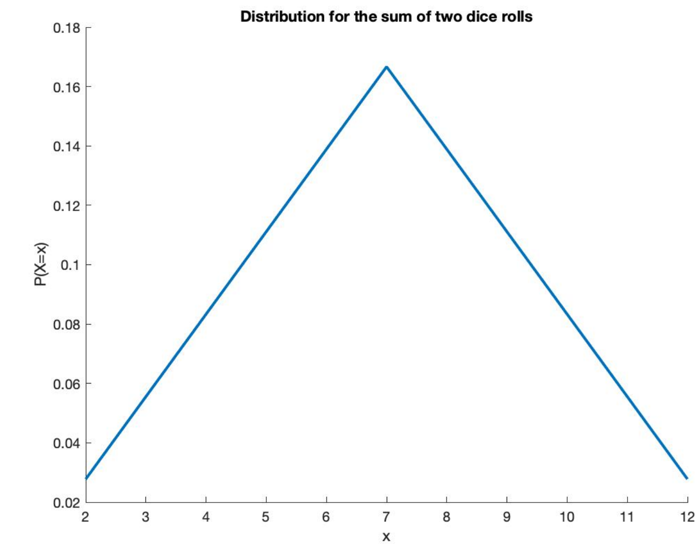
A discrete random variable is one that can equal a **finite number of values**. We can describe the distribution of a discrete r.v. using a **probability mass function (pmf)**, which must be non-negative for all inputs and sum to 1. A pmf describes the probability that a discrete r.v. X takes on a particular value, $P(X = x)$.

What is the pmf of (1)?

Examples of discrete r.v.'s

(1) X = sum of two rolled dice
(2) X = maximum of two rolled dice
(3) X = # of coin flips until the 1st heads
(4) X = number of kids in a family

DISCRETE PROBABILITY DISTRIBUTIONS: PMF



A discrete random variable is one that can equal a **finite number of values**. We can describe the distribution of a discrete r.v. using a **probability mass function (pmf)**, which must be non-negative for all inputs and sum to 1. A pmf describes the probability that a discrete r.v. X takes on a particular value, $P(X = x)$.

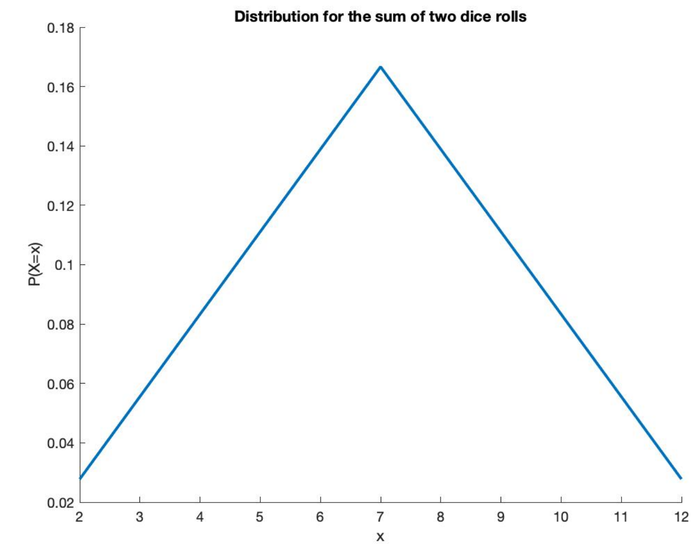
Examples of discrete r.v.'s

(1) X = sum of two rolled dice
(2) X = maximum of two rolled dice
(3) X = # of coin flips until the 1st heads
(4) X = number of kids in a family

What is the pmf of (1)?

x	2	3	4	5	6	7	8	9	10	11	12
P(X=x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

DISCRETE PROBABILITY DISTRIBUTIONS: PMF



A discrete random variable is one that can equal a **finite number of values**. We can describe the distribution of a discrete r.v. using a **probability mass function (pmf)**, which must be non-negative for all inputs and sum to 1. A pmf describes the probability that a discrete r.v. X takes on a particular value, $P(X = x)$.

What is the pmf of (1)?

x	2	3	4	5	6	7	8	9	10	11	12
P(X=x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Now, what is the probability that the sum of two rolls is at least 5? What about no more than 8?

Examples of discrete r.v.'s

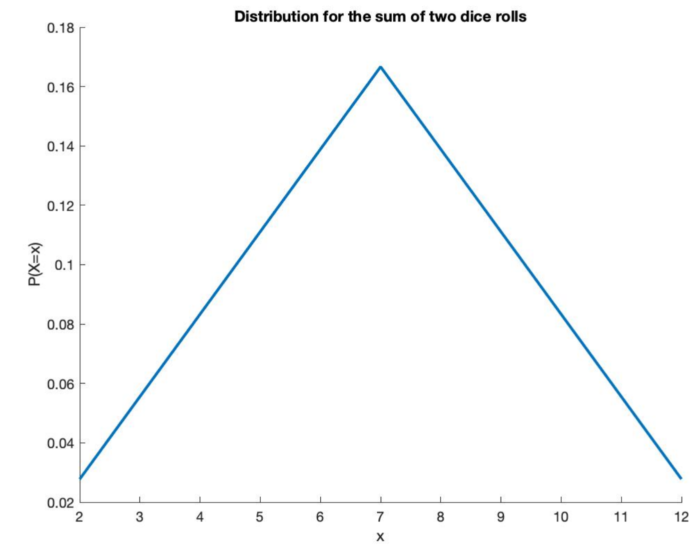
(1) X = sum of two rolled dice

(2) X = maximum of two rolled dice

(3) X = # of coin flips until the 1st heads

(4) X = number of kids in a family

DISCRETE PROBABILITY DISTRIBUTIONS: PMF



A discrete random variable is one that can equal a **finite number of values**. We can describe the distribution of a discrete r.v. using a **probability mass function (pmf)**, which must be non-negative for all inputs and sum to 1. A pmf describes the probability that a discrete r.v. X takes on a particular value, $P(X = x)$.

Examples of discrete r.v.'s

(1) X = sum of two rolled dice

(2) X = maximum of two rolled dice

(3) X = # of coin flips until the 1st heads

(4) X = number of kids in a family

What is the pmf of (1)?

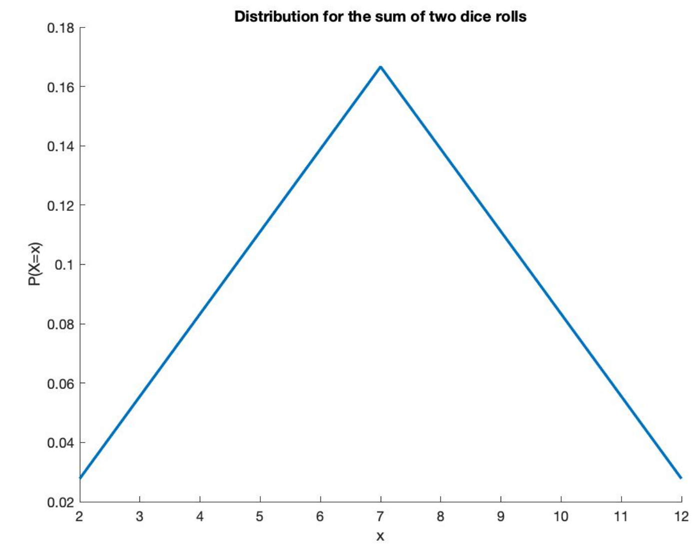
x	2	3	4	5	6	7	8	9	10	11	12
P(X=x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Now, what is the probability that the sum of two rolls is at least 5? What about no more than 8?

$$P(X \geq 5) = \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{25}{36}$$

DISCRETE PROBABILITY DISTRIBUTIONS: PMF

A discrete random variable is one that can equal a **finite number of values**. We can describe the distribution of a discrete r.v. using a **probability mass function (pmf)**, which must be non-negative for all inputs and sum to 1. A pmf describes the probability that a discrete r.v. X takes on a particular value, $P(X = x)$.



Examples of discrete r.v.'s

(1) X = sum of two rolled dice

(2) X = maximum of two rolled dice

(3) X = # of coin flips until the 1st heads

(4) X = number of kids in a family

What is the pmf of (1)?

x	2	3	4	5	6	7	8	9	10	11	12
P(X=x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Now, what is the probability that the sum of two rolls is at least 5? What about no more than 8?

$$P(X \geq 5) = \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{25}{36}$$

$$P(X \leq 8) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} = \frac{26}{36}$$

DISCRETE PROBABILITY DISTRIBUTIONS: CDF

Another way to describe a r.v.'s distribution is with a **cumulative distribution function (cdf)**, which is a non-decreasing, right continuous step function for discrete r.v.'s. A cdf describes the probability that a discrete r.v. X is less than or equal to a particular value k, $P(X \leq k)$.

DISCRETE PROBABILITY DISTRIBUTIONS: CDF

Another way to describe a r.v.'s distribution is with a **cumulative distribution function (cdf)**, which is a non-decreasing, right continuous step function for discrete r.v.'s. A cdf describes the probability that a discrete r.v. X is less than or equal to a particular value k, $P(X \leq k)$.

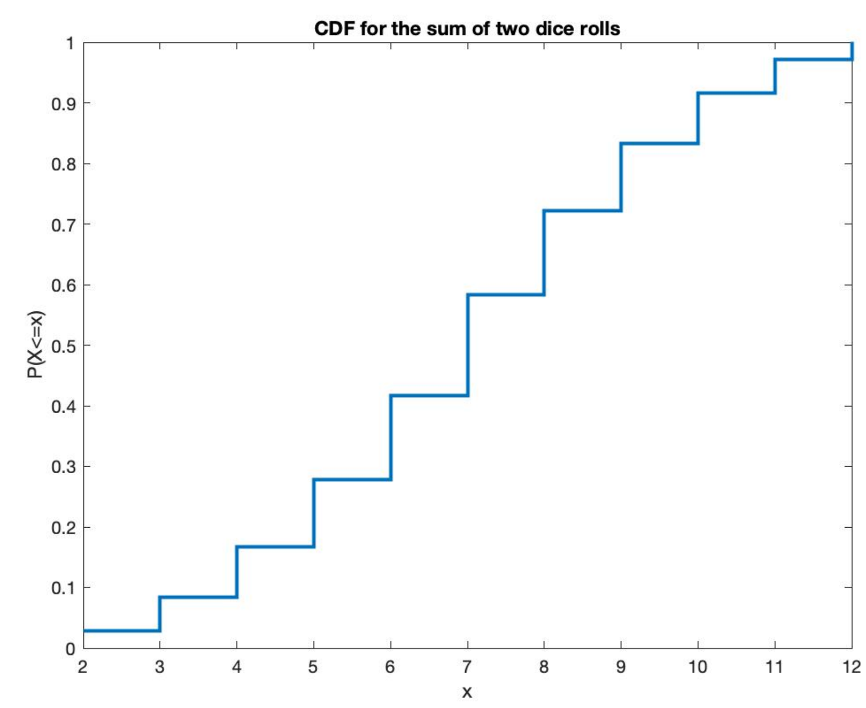
Back to our example: what is the cdf of the sum of two dice rolls?

DISCRETE PROBABILITY DISTRIBUTIONS: CDF

Another way to describe a r.v.'s distribution is with a **cumulative distribution function (cdf)**, which is a non-decreasing, right continuous step function for discrete r.v.'s. A cdf describes the probability that a discrete r.v. X is less than or equal to a particular value k, $P(X \leq k)$.

Back to our example: what is the cdf of the sum of two dice rolls?

x	2	3	4	5	6	7	8	9	10	11	12
P(X≤x)	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36



DISCRETE PROBABILITY DISTRIBUTIONS: EXAMPLES

Binomial Distribution

The distribution for the number of successes in a sequence of n independent experiments, each with 2 possible outcomes where p is the probability of success. Example: flipping a coin.

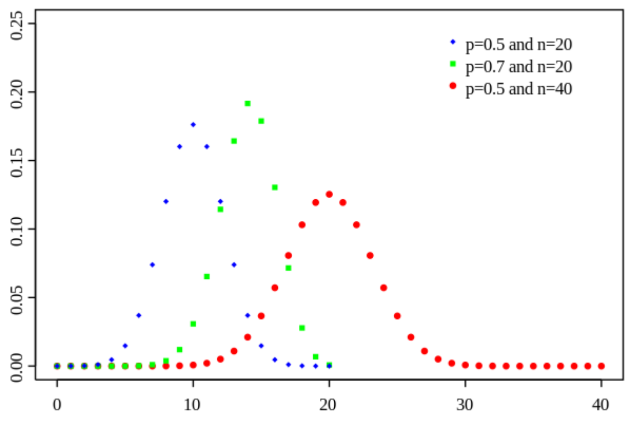
$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

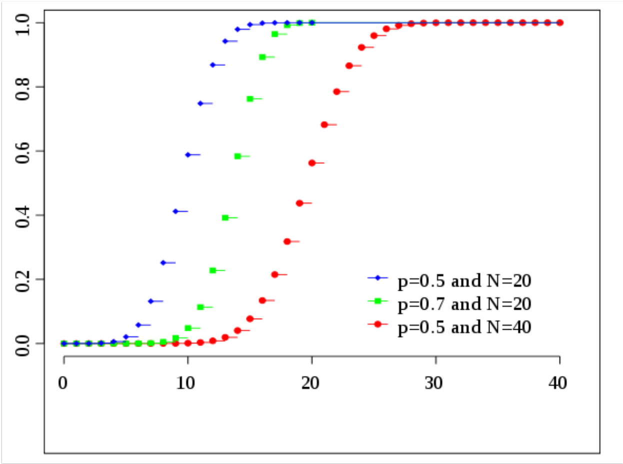
DISCRETE PROBABILITY DISTRIBUTIONS: EXAMPLES

Binomial Distribution

The distribution for the number of successes in a sequence of n independent experiments, each with 2 possible outcomes where p is the probability of success. Example: flipping a coin.

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



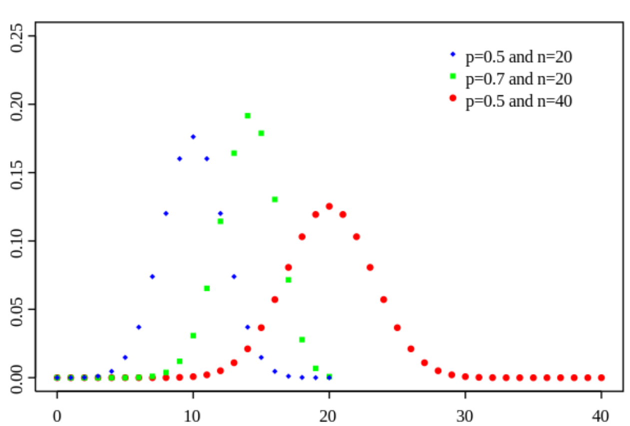


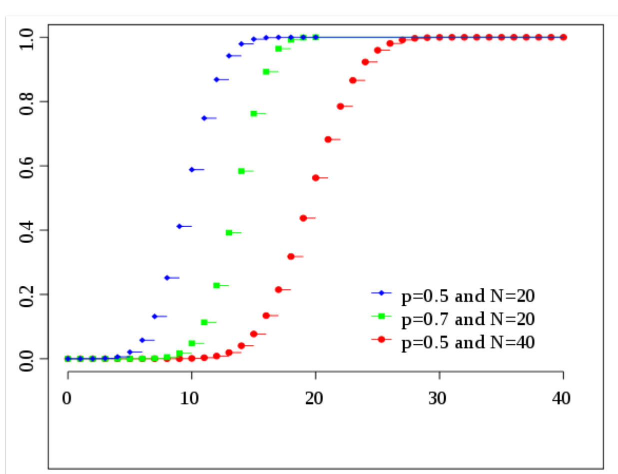
DISCRETE PROBABILITY DISTRIBUTIONS: EXAMPLES

Binomial Distribution

The distribution for the number of successes in a sequence of n independent experiments, each with 2 possible outcomes where p is the probability of success. Example: flipping a coin.

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





Poisson Distribution

The distribution that expresses the probability that a given number of events will occur in a fixed interval of time λ. The event must occur with a constant rate and independently of the last event. Example: neural spike counts.

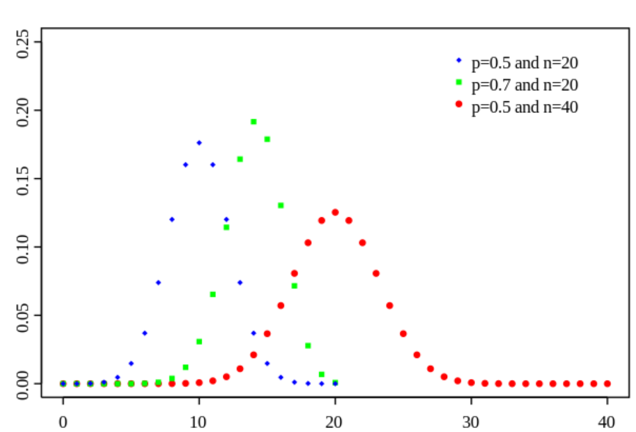
$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

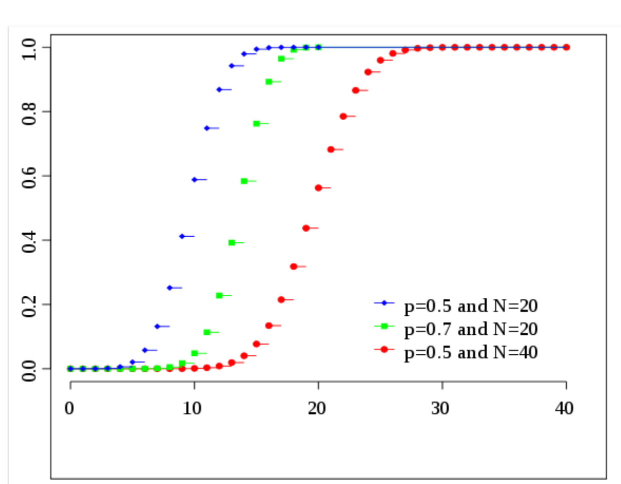
DISCRETE PROBABILITY DISTRIBUTIONS: EXAMPLES

Binomial Distribution

The distribution for the number of successes in a sequence of n independent experiments, each with 2 possible outcomes where p is the probability of success. Example: flipping a coin.

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

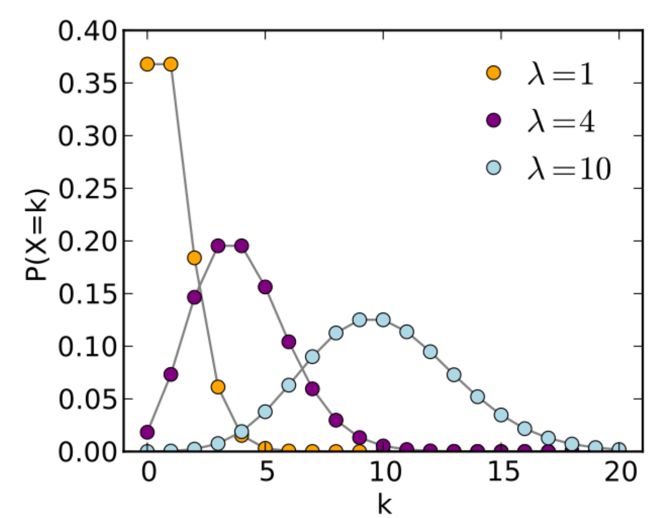


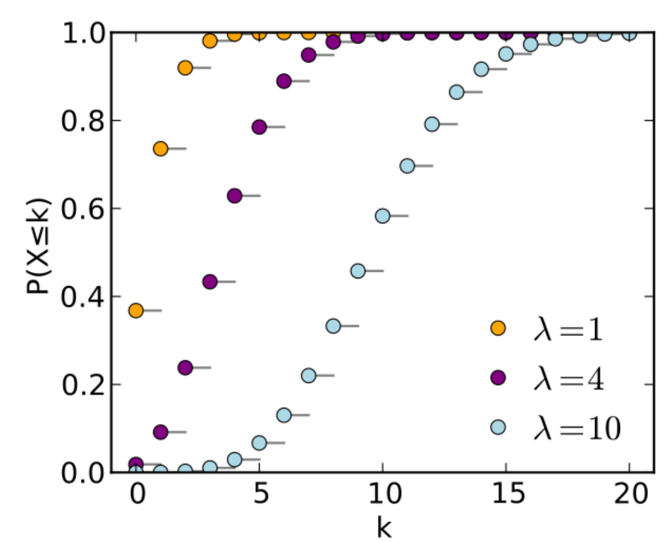


Poisson Distribution

The distribution that expresses the probability that a given number of events will occur in a fixed interval of time λ. The event must occur with a constant rate and independently of the last event. Example: neural spike counts.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$





CONTINUOUS PROBABILITY DISTRIBUTIONS: PDF AND CDF

Of course, we sometimes have r.v.'s whose possible values can't be listed because they are **continuous**. In this case, we use a smooth curve called a **probability density function (pdf)** to assign a distribution to the r.v. To be a valid pdf, $f(x)$ must be non-negative and integrate to 1. The **area under the pdf is what represents probability**. Our cdf is still defined as $P(X \leq k)$, but this is now the integral of the pdf $f_x(x)$.

CONTINUOUS PROBABILITY DISTRIBUTIONS: PDF AND CDF

Of course, we sometimes have r.v.'s whose possible values can't be listed because they are **continuous**. In this case, we use a smooth curve called a **probability density function (pdf)** to assign a distribution to the r.v. To be a valid pdf, $f(x)$ must be non-negative and integrate to 1. The **area under the pdf is what represents probability**. Our cdf is still defined as $P(X \leq k)$, but this is now the integral of the pdf $f_x(x)$.

Example: calculating probability using a pdf

Suppose $f(y) = 4y^3$ for $0 < y < 1$. Find $P(0 < Y < 0.5)$.

CONTINUOUS PROBABILITY DISTRIBUTIONS: PDF AND CDF

Of course, we sometimes have r.v.'s whose possible values can't be listed because they are **continuous**. In this case, we use a smooth curve called a **probability density function (pdf)** to assign a distribution to the r.v. To be a valid pdf, $f(x)$ must be non-negative and integrate to 1. The **area under the pdf is what represents probability**. Our cdf is still defined as $P(X \leq k)$, but this is now the integral of the pdf $f_x(x)$.

Example: calculating probability using a pdf

Suppose $f(y) = 4y^3$ for $0 < y < 1$. Find $P(0 < Y < 0.5)$.

$$P(0 < Y < 0.5) = \int_0^{0.5} 4y^3 dy = y^4 |_0^{0.5} = 0.0625$$

CONTINUOUS PROBABILITY DISTRIBUTIONS: PDF AND CDF

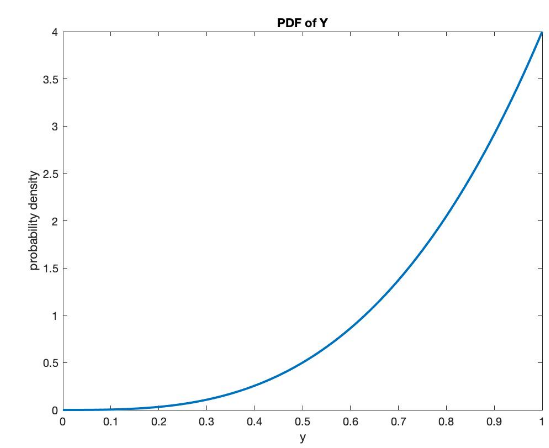
Of course, we sometimes have r.v.'s whose possible values can't be listed because they are **continuous**. In this case, we use a smooth curve called a **probability density function (pdf)** to assign a distribution to the r.v. To be a valid pdf, $f(x)$ must be non-negative and integrate to 1. The **area under the pdf is what represents probability**. Our cdf is still defined as $P(X \leq k)$, but this is now the integral of the pdf $f_x(x)$.

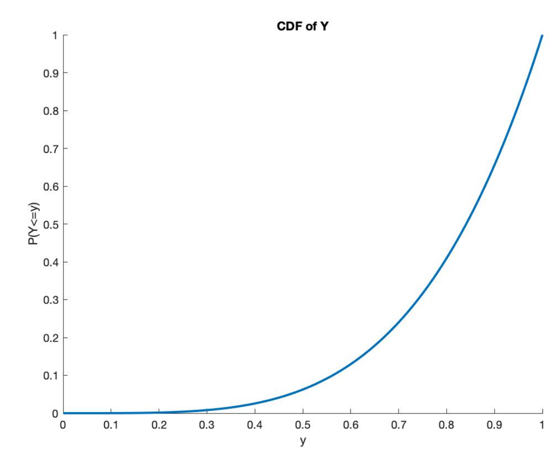
Example: calculating probability using a pdf

Suppose $f(y) = 4y^3$ for $0 < y < 1$. Find $P(0 < Y < 0.5)$.

$$P(0 < Y < 0.5) = \int_0^{0.5} 4y^3 dy = y^4 \,|_0^{0.5} = 0.0625$$

Note: the y axis of a continuous pdf no longer represents probability as in the discrete case, but rather a probability density. Additionally, the probability for a single x value occurring is always 0 and instead we typically look at a range of x values and integrate to find probability.





CONTINUOUS PROBABILITY DISTRIBUTIONS: EXAMPLES

Exponential Distribution

The distribution for the time between events that occur continuously and independently at a constant average rate $\lambda = \frac{1}{\beta}$ (this is a specific case of the gamma distribution). Example: time spent waiting at a restaurant before being served.

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

The exponential distribution is memoryless, meaning that the past has no bearing on its future behavior.

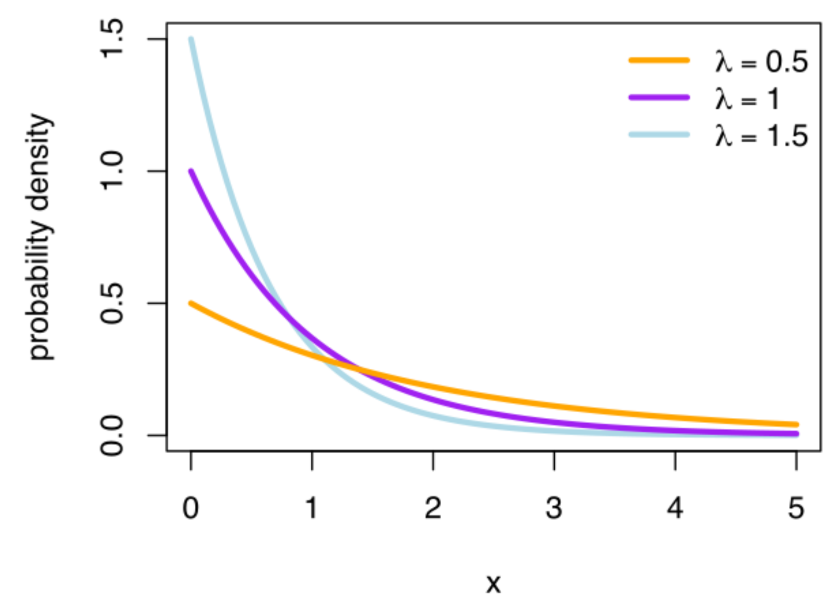
CONTINUOUS PROBABILITY DISTRIBUTIONS: EXAMPLES

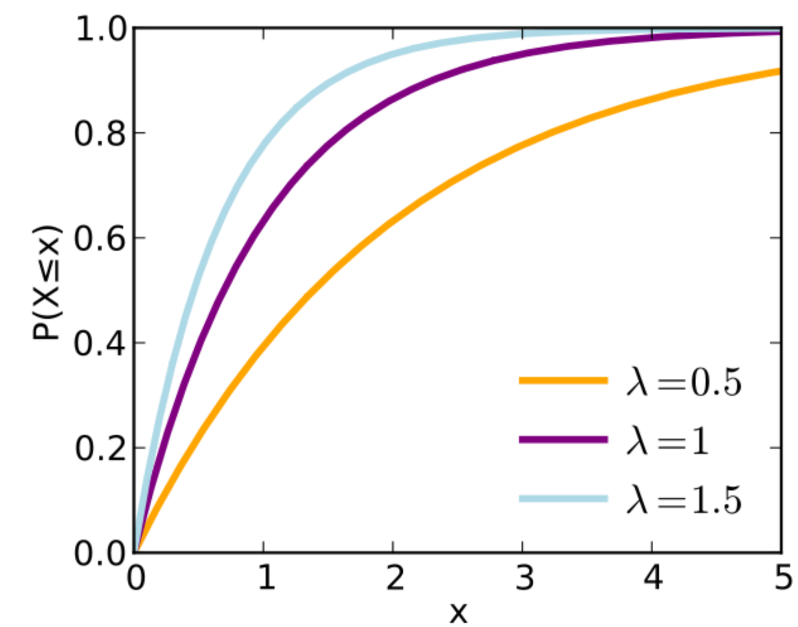
Exponential Distribution

The distribution for the time between events that occur continuously and independently at a constant average rate $\lambda = \frac{1}{\beta}$ (this is a specific case of the gamma distribution). Example: time spent waiting at a restaurant before being served.

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

The exponential distribution is memoryless, meaning that the past has no bearing on its future behavior.





EXPECTED VALUE

The **expected value** of a r.v. X is a **weighted average of all the possible values** of X.

Discrete r.v.: $E(X)$ is a sum

Continuous r.v.: $E(X)$ is an integral

Linearity of expectation: If $Y = aX + b$, then $E(Y) = E(aX + b) = aE(X) + b$

EXPECTED VALUE

The **expected value** of a r.v. X is a **weighted average of all the possible values** of X.

Discrete r.v.: $E(X)$ is a sum

Continuous r.v.: $E(X)$ is an integral

Linearity of expectation: If $Y = aX + b$, then $E(Y) = E(aX + b) = aE(X) + b$

Example 1: Suppose you roll a die, and are paid \$1 for odd rolls and \$2 for even rolls. What is the expected value for one roll?

EXPECTED VALUE

The **expected value** of a r.v. X is a **weighted average of all the possible values** of X.

Discrete r.v.: $E(X)$ is a sum

Continuous r.v.: $E(X)$ is an integral

Linearity of expectation: If $Y = aX + b$, then $E(Y) = E(aX + b) = aE(X) + b$

Example 1: Suppose you roll a die, and are paid $1 for odd rolls and $2 for even rolls. What is the expected value for one roll?

x	1	2	3	4	5	6
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6
Amount ($)	1	2	1	2	1	2

$$E(X) = 1(\frac{1}{6}) + 2(\frac{1}{6}) + 1(\frac{1}{6}) + 2(\frac{1}{6}) + 1(\frac{1}{6}) + 2(\frac{1}{6}) = \frac{9}{6} = 1.5$$

EXPECTED VALUE

The **expected value** of a r.v. X is a **weighted average of all the possible values** of X.

Discrete r.v.: $E(X)$ is a sum
Continuous r.v.: $E(X)$ is an integral
Linearity of expectation: If $Y = aX + b$, then $E(Y) = E(aX + b) = aE(X) + b$

Example 1: Suppose you roll a die, and are paid $1 for odd rolls and $2 for even rolls. What is the expected value for one roll?

x	1	2	3	4	5	6
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6
Amount ($)	1	2	1	2	1	2

$$E(X) = 1(\frac{1}{6}) + 2(\frac{1}{6}) + 1(\frac{1}{6}) + 2(\frac{1}{6}) + 1(\frac{1}{6}) + 2(\frac{1}{6}) = \frac{9}{6} = 1.5$$

Example 2: St. Petersburg Paradox
A fair coin is flipped until the 1st tail appears, and you win $\$2^k$ if it appears on the k^{th} toss. If X = your winnings, how much should you pay in order for this to be a fair game?

EXPECTED VALUE

The **expected value** of a r.v. X is a **weighted average of all the possible values** of X.

Discrete r.v.: $E(X)$ is a sum
Continuous r.v.: $E(X)$ is an integral
Linearity of expectation: If $Y = aX + b$, then $E(Y) = E(aX + b) = aE(X) + b$

Example 1: Suppose you roll a die, and are paid \$1 for odd rolls and \$2 for even rolls. What is the expected value for one roll?

x	1	2	3	4	5	6
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6
Amount (\$)	1	2	1	2	1	2

$$E(X) = 1(\frac{1}{6}) + 2(\frac{1}{6}) + 1(\frac{1}{6}) + 2(\frac{1}{6}) + 1(\frac{1}{6}) + 2(\frac{1}{6}) = \frac{9}{6} = 1.5$$

Example 2: St. Petersburg Paradox
A fair coin is flipped until the 1st tail appears, and you win \$$2^k$ if it appears on the k^{th} toss. If X = your winnings, how much should you pay in order for this to be a fair game?

$$E(X) = \sum kP(x = k) = 2(\frac{1}{2}) + 4(\frac{1}{4}) + \ldots = \infty$$

This leads to an ill-advised gambling strategy where you are guaranteed to make \$1 (or an arbitrarily large amount of money)

VARIANCE AND STANDARD DEVIATION

For any r.v., we want to summarize its central tendency but also its spread. The variance of a r.v. X is given by finding the **average squared deviation** of X from its mean, $E(X)$.

$$Var(X) = E[(X - E(X))^2] \text{ or equivalently } Var(X) = E(X^2) - [E(X)]^2$$

VARIANCE AND STANDARD DEVIATION

For any r.v., we want to summarize its central tendency but also its spread. The variance of a r.v. X is given by finding the **average squared deviation** of X from its mean, $E(X)$.

$$Var(X) = E[(X - E(X))^2] \text{ or equivalently } Var(X) = E(X^2) - [E(X)]^2$$

Since variance averages **squared deviations**, its units are the square of the units of the original r.v. So we define **standard deviation**:

$$sd(X) = sqrt(Var(X))$$

VARIANCE AND STANDARD DEVIATION

For any r.v., we want to summarize its central tendency but also its spread. The variance of a r.v. X is given by finding the **average squared deviation** of X from its mean, $E(X)$.

$$Var(X) = E[(X - E(X))^2] \text{ or equivalently } Var(X) = E(X^2) - [E(X)]^2$$

Since variance averages **squared deviations**, its units are the square of the units of the original r.v. So we define **standard deviation**:

$$sd(X) = sqrt(Var(X))$$

Example: Let X = the number of bases a baseball player earns per at-bat. Given the probability function below, find the expected value, variance, and standard deviation of X.

k	0	1	2	3	4
P(X=k)	0.65	0.25	0.06	0.01	0.03

VARIANCE AND STANDARD DEVIATION

For any r.v., we want to summarize its central tendency but also its spread. The variance of a r.v. X is given by finding the **average squared deviation** of X from its mean, $E(X)$.

$$Var(X) = E[(X - E(X))^2] \text{ or equivalently } Var(X) = E(X^2) - [E(X)]^2$$

Since variance averages **squared deviations**, its units are the square of the units of the original r.v. So we define **standard deviation**:

$$sd(X) = sqrt(Var(X))$$

Example: Let X = the number of bases a baseball player earns per at-bat. Given the probability function below, find the expected value, variance, and standard deviation of X.

$$E(X) = 0(0.65) + 1(0.25) + 2(0.06) + 3(0.01) + 4(0.03) = 0.52$$

$$E(X^2) = \sum_{k=0}^{4} k^2 P(X = k) = 1.06$$

k	0	1	2	3	4
P(X=k)	0.65	0.25	0.06	0.01	0.03

VARIANCE AND STANDARD DEVIATION

For any r.v., we want to summarize its central tendency but also its spread. The variance of a r.v. X is given by finding the **average squared deviation** of X from its mean, $E(X)$.

$$Var(X) = E[(X - E(X))^2] \text{ or equivalently } Var(X) = E(X^2) - [E(X)]^2$$

Since variance averages **squared deviations**, its units are the square of the units of the original r.v. So we define **standard deviation**:

$$sd(X) = sqrt(Var(X))$$

Example: Let X = the number of bases a baseball player earns per at-bat. Given the probability function below, find the expected value, variance, and standard deviation of X.

k	0	1	2	3	4
P(X=k)	0.65	0.25	0.06	0.01	0.03

$$E(X) = 0(0.65) + 1(0.25) + 2(0.06) + 3(0.01) + 4(0.03) = 0.52$$

$$E(X^2) = \sum_{k=0}^{4} k^2 P(X = k) = 1.06$$

$$Var(X) = E(X^2) - E(X)^2 = 1.06 - 0.52^2 = 0.7896$$

$$sd(X) = \sqrt{Var(X)} = 0.8886$$

MOMENTS

If X is a r.v., then we can define $E(X^n)$ as the n^{th} **moment of** X. Each moment gives us some insight into the characteristics of the distribution.

In the discrete case, $E(X^n) = \sum_i x_i^n p(x)$ or centered $\sum_i (x_i - \mu)^n p(x_i)$

In the continuous case, $E(X^n) = \int_{-\infty}^{\infty} x^n f(x)dx$ or centered $\int_{-\infty}^{\infty} (x - \mu)^n f(x)dx$

MOMENTS

If X is a r.v., then we can define $E(X^n)$ as the n^{th} **moment of X**. Each moment gives us some insight into the characteristics of the distribution.

In the discrete case, $E(X^n) = \sum_i x_i^n p(x)$ or centered $\sum_i (x_i - \mu)^n p(x_i)$

In the continuous case, $E(X^n) = \int_{-\infty}^{\infty} x^n f(x) dx$ or centered $\int_{-\infty}^{\infty} (x - \mu)^n f(x) dx$

First moment: mean

Central tendency, ie the sum of the products and their probabilities (their average)

MOMENTS

If X is a r.v., then we can define $E(X^n)$ as the n^{th} **moment of** X. Each moment gives us some insight into the characteristics of the distribution.

In the discrete case, $E(X^n) = \sum_i x_i^n p(x)$ or centered $\sum_i (x_i - \mu)^n p(x_i)$

In the continuous case, $E(X^n) = \int_{-\infty}^{\infty} x^n f(x) dx$ or centered $\int_{-\infty}^{\infty} (x - \mu)^n f(x) dx$

First moment: mean

Central tendency, ie the sum of the products and their probabilities (their average)

Second moment: variance

Spread of the observations from their average value, ie the squared deviation of the r.v. from the mean

MOMENTS

If X is a r.v., then we can define $E(X^n)$ as the n^{th} **moment of X**. Each moment gives us some insight into the characteristics of the distribution.

In the discrete case, $E(X^n) = \sum_i x_i^n p(x)$ or centered $\sum_i (x_i - \mu)^n p(x_i)$

In the continuous case, $E(X^n) = \int_{-\infty}^{\infty} x^n f(x) dx$ or centered $\int_{-\infty}^{\infty} (x - \mu)^n f(x) dx$

First moment: mean

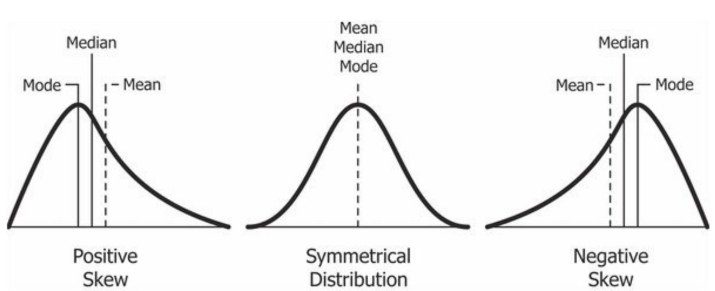
Central tendency, ie the sum of the products and their probabilities (their average)

Second moment: variance

Spread of the observations from their average value, ie the squared deviation of the r.v. from the mean

Third moment: skewness

Symmetry of the distribution around the mean



MOMENTS

If X is a r.v., then we can define $E(X^n)$ as the n^{th} **moment of X**. Each moment gives us some insight into the characteristics of the distribution.

In the discrete case, $E(X^n) = \sum_i x_i^n p(x)$ or centered $\sum_i (x_i - \mu)^n p(x_i)$

In the continuous case, $E(X^n) = \int_{-\infty}^{\infty} x^n f(x)dx$ or centered $\int_{-\infty}^{\infty} (x-\mu)^n f(x)dx$

First moment: mean

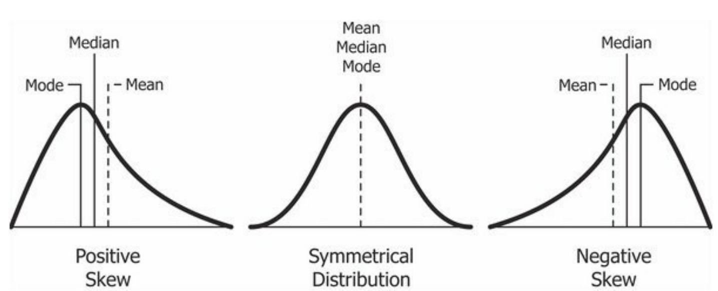
Central tendency, ie the sum of the products and their probabilities (their average)

Second moment: variance

Spread of the observations from their average value, ie the squared deviation of the r.v. from the mean

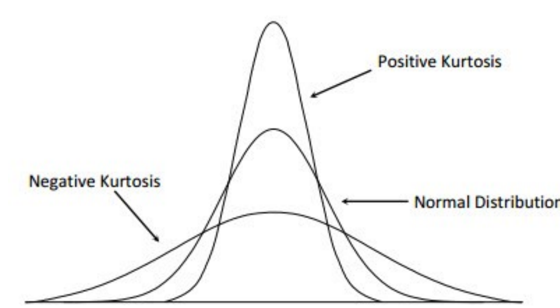
Third moment: skewness

Symmetry of the distribution around the mean



Fourth moment: kurtosis

How heavy the tails of the distribution are



CODE BREAK 2

COVARIANCE AND CORRELATION

Covariance measures the **joint variability of two r.v.'s**. If both X and Y tend to be "big" at the same time, then $Cov(X, Y) > 0$. If one tends to be "big" when the other is "small", then $Cov(X, Y) < 0$.

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$
$$Cov(X, Y) = E(XY) - E(X)E(Y)$$
$$Cov(X, X) = Var(X)$$

COVARIANCE AND CORRELATION

Covariance measures the **joint variability of two r.v.'s**. If both X and Y tend to be "big" at the same time, then $Cov(X, Y) > 0$. If one tends to be "big" when the other is "small", then $Cov(X, Y) < 0$.

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$
$$Cov(X, Y) = E(XY) - E(X)E(Y)$$
$$Cov(X, X) = Var(X)$$

Correlation measures the **linear relation between two r.v.'s**, and is just the covariance of the variables divided by the product of their standard deviations.

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$
$$-1 \leq Corr(X, Y) \leq 1$$

COVARIANCE AND CORRELATION

Covariance measures the **joint variability of two r.v.'s**. If both X and Y tend to be "big" at the same time, then $Cov(X, Y) > 0$. If one tends to be "big" when the other is "small", then $Cov(X, Y) < 0$.

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$
$$Cov(X, Y) = E(XY) - E(X)E(Y)$$
$$Cov(X, X) = Var(X)$$

Example: Suppose X and Y are discrete r.v.'s with the joint probability function below. Find $Corr(X, Y)$.

(x,y)	(1,2)	(1,3)	(2,1)	(2,4)
P(X=x, Y=y)	0.5	0.25	0.125	0.125

Correlation measures the **linear relation between two r.v.'s**, and is just the covariance of the variables divided by the product of their standard deviations.

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$
$$-1 \leq Corr(X, Y) \leq 1$$

COVARIANCE AND CORRELATION

Covariance measures the **joint variability of two r.v.'s**. If both X and Y tend to be "big" at the same time, then $Cov(X, Y) > 0$. If one tends to be "big" when the other is "small", then $Cov(X, Y) < 0$.

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$
$$Cov(X, Y) = E(XY) - E(X)E(Y)$$
$$Cov(X, X) = Var(X)$$

Example: Suppose X and Y are discrete r.v.'s with the joint probability function below. Find $Corr(X, Y)$.

(x,y)	(1,2)	(1,3)	(2,1)	(2,4)
P(X=x, Y=y)	0.5	0.25	0.125	0.125

Correlation measures the **linear relation between two r.v.'s**, and is just the covariance of the variables divided by the product of their standard deviations.

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$
$$-1 \leq Corr(X, Y) \leq 1$$

$$E(X) = 1(0.75) + 2(0.25) = 1.25$$
$$E(Y) = 2(0.5) + 3(0.25) + 1(0.125) + 4(0.125) = 2.375$$
$$E(XY) = 2(0.625) + 3(0.25) + 8(0.125) = 3$$

COVARIANCE AND CORRELATION

Covariance measures the **joint variability of two r.v.'s**. If both X and Y tend to be "big" at the same time, then $Cov(X, Y) > 0$. If one tends to be "big" when the other is "small", then $Cov(X, Y) < 0$.

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$
$$Cov(X, Y) = E(XY) - E(X)E(Y)$$
$$Cov(X, X) = Var(X)$$

Example: Suppose X and Y are discrete r.v.'s with the joint probability function below. Find $Corr(X, Y)$.

(x,y)	(1,2)	(1,3)	(2,1)	(2,4)
P(X=x, Y=y)	0.5	0.25	0.125	0.125

Correlation measures the **linear relation between two r.v.'s**, and is just the covariance of the variables divided by the product of their standard deviations.

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$
$$-1 \leq Corr(X, Y) \leq 1$$

$E(X) = 1(0.75) + 2(0.25) = 1.25$

$E(Y) = 2(0.5) + 3(0.25) + 1(0.125) + 4(0.125) = 2.375$

$E(XY) = 2(0.625) + 3(0.25) + 8(0.125) = 3$

$E(X^2) = 1.75, Var(X) = 0.1875$

$E(Y^2) = 6.375, Var(Y) = 0.1734$

COVARIANCE AND CORRELATION

Covariance measures the **joint variability of two r.v.'s**. If both X and Y tend to be "big" at the same time, then $Cov(X, Y) > 0$. If one tends to be "big" when the other is "small", then $Cov(X, Y) < 0$.

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$
$$Cov(X, Y) = E(XY) - E(X)E(Y)$$
$$Cov(X, X) = Var(X)$$

Example: Suppose X and Y are discrete r.v.'s with the joint probability function below. Find $Corr(X, Y)$.

(x,y)	(1,2)	(1,3)	(2,1)	(2,4)
P(X=x, Y=y)	0.5	0.25	0.125	0.125

Correlation measures the **linear relation between two r.v.'s**, and is just the covariance of the variables divided by the product of their standard deviations.

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$
$$-1 \leq Corr(X, Y) \leq 1$$

$$E(X) = 1(0.75) + 2(0.25) = 1.25$$
$$E(Y) = 2(0.5) + 3(0.25) + 1(0.125) + 4(0.125) = 2.375$$
$$E(XY) = 2(0.625) + 3(0.25) + 8(0.125) = 3$$
$$E(X^2) = 1.75, Var(X) = 0.1875$$
$$E(Y^2) = 6.375, Var(Y) = 0.1734$$
$$Cov(X, Y) = E(XY) - E(X)E(Y) = 0.03125$$
$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = 0.0842$$

THE COVARIANCE MATRIX

The covariance matrix is a matrix whose element in the i, j position represents **the covariance between the i^{th} and j^{th} element**s of a random vector (a r.v. with multiple dimensions). This generalizes the notion of covariance to multiple dimensions. For example, the variation of a collection of two-dimensional points cannot be fully characterized by a single number or just the variances in the x and y directions: a $2x2$ matrix is needed.

THE COVARIANCE MATRIX

The covariance matrix is a matrix whose element in the i, j position represents **the covariance between the i^{th} and j^{th} elements** of a random vector (a r.v. with multiple dimensions). This generalizes the notion of covariance to multiple dimensions. For example, the variation of a collection of two-dimensional points cannot be fully characterized by a single number or just the variances in the x and y directions: a $2x2$ matrix is needed.

Assume we have a matrix $X = (X_1, X_2, \ldots, X_n)^T$ where each entry X_i is the entire distribution of a random variable.

We each entry of the covariance matrix, K_{xx}, as:

$K_{X_iX_j} = Cov(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))]$

THE COVARIANCE MATRIX

The covariance matrix is a matrix whose element in the i, j position represents **the covariance between the i^{th} and j^{th} elements** of a random vector (a r.v. with multiple dimensions). This generalizes the notion of covariance to multiple dimensions. For example, the variation of a collection of two-dimensional points cannot be fully characterized by a single number or just the variances in the x and y directions: a $2x2$ matrix is needed.

Assume we have a matrix $X = (X_1, X_2, \ldots, X_n)^T$ where each entry X_i is the entire distribution of a random variable.
We each entry of the covariance matrix, K_{xx}, as:
$K_{X_iX_j} = Cov(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))]$
So the full matrix for n r.v.'s is equal to the matrix equation $E(XX^T) - \mu_x\mu_x^T$:

$$\mathrm{K}_{\mathbf{XX}} = \begin{bmatrix} \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_1 - \mathrm{E}[X_1])] & \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_2 - \mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_n - \mathrm{E}[X_n])] \\ \mathrm{E}[(X_2 - \mathrm{E}[X_2])(X_1 - \mathrm{E}[X_1])] & \mathrm{E}[(X_2 - \mathrm{E}[X_2])(X_2 - \mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_2 - \mathrm{E}[X_2])(X_n - \mathrm{E}[X_n])] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mathrm{E}[X_n])(X_1 - \mathrm{E}[X_1])] & \mathrm{E}[(X_n - \mathrm{E}[X_n])(X_2 - \mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_n - \mathrm{E}[X_n])(X_n - \mathrm{E}[X_n])] \end{bmatrix}$$

CODE BREAK 3