# SML 312 – Probability and Statistics Review

Course Overview

# Outline

1. Basics of Probability – Distributions and their properties/meaning

# Outline

1. Basics of Probability – Distributions and their properties/meaning
2. Numbers computed from probabilities: Expected Values, Mean, Variance, Covariance

# Outline

1. Basics of Probability – Distributions and their properties/meaning
2. Numbers computed from probabilities: Expected Values, Mean, Variance, Covariance
3. Statistics – Sample mean, sample variance, estimators

# Outline

1. Basics of Probability – Distributions and their properties/meaning
2. Numbers computed from probabilities: Expected Values, Mean, Variance, Covariance
3. Statistics – Sample mean, sample variance, estimators
4. Role in Data Science

# Probability Review

**Probability** is the mathematical study of events and their outcomes (weighted by probabilities). There can be finitely many or infinitely many possible outcomes. The mathematics of probability may look different in these two cases, but all ideas are the same. A **random variable (RV)** X drawn from a given probability distribution P is denoted as X ~ P. We often speak of random variables instead of probability distributions.

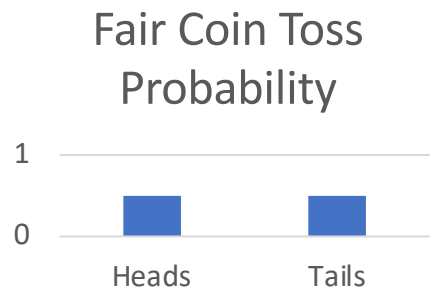
# Probability Review

**Probability** is the mathematical study of events and their outcomes (weighted by probabilities). There can be finitely many or infinitely many possible outcomes. The mathematics of probability may look different in these two cases, but all ideas are the same. A **random variable (RV)** X drawn from a given probability distribution P is denoted as X ~ P. We often speak of random variables instead of probability distributions.

1) Finitely many outcomes (**discrete** RV): Finite sums, event probabilities are non-negative numbers. The sum of all (disjoint) event probabilities is 1.

Example: Fair Coin Toss:

$$P(\text{heads}) = \tfrac{1}{2} \text{ and } P(\text{tails}) = \tfrac{1}{2}$$



# Probability Review

**Probability** is the mathematical study of events and their outcomes (weighted by probabilities). There can be finitely many or infinitely many possible outcomes. The mathematics of probability may look different in these two cases, but all ideas are the same. A **random variable (RV)** X drawn from a given probability distribution P is denoted as X ~ P. We often speak of random variables instead of probability distributions.

1) Finitely many outcomes (**discrete** RV): Finite sums, event probabilities are non-negative numbers. The sum of all (disjoint) event probabilities is 1.

Example: Fair Coin Toss:

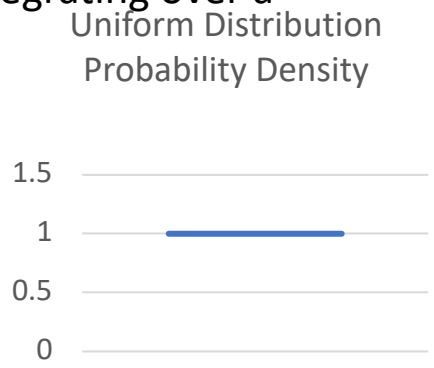
P(heads) = ½ and P(tails) = ½


2) Infinitely many outcomes (**continuous** RV): Sums are replaced by integrals, event probabilities are described by a non-negative probability distribution function whose total integral is 1. Numerical probabilities are assigned to events by integrating over a set of possible outcomes.

Example: Uniform distribution on [0,1]

This has **probability distribution function** p(x) given by

p(x) = 1 when 0 <= x <= 1 and p(x) = 0 otherwise



# Probability Review

Suppose that X is a random variable with probability distribution P(x) or probability density p(x). Then we define:

Expected value =

# Probability Review

Suppose that X is a random variable with probability distribution P(x) or probability density p(x). Then we define:

Expected value = $\mathbb{E}[X]$

# Probability Review

Suppose that X is a random variable with probability distribution P(x) or probability density p(x). Then we define:

$$\text{Expected value} = \mathbb{E}[X] \coloneqq \begin{cases} \sum_{i \epsilon I} P(x_i) * x_i \,, & if\ X\ is\ discrete \\ \int_{x \epsilon \mathbb{R}} p(x) * x\ * dx \,, & if\ X\ is\ continuous \end{cases}$$

# Probability Review

Suppose that X is a random variable with probability distribution P(x) or probability density p(x). Then we define:

$$\text{Expected value} = \mathbb{E}[X] := \begin{cases} \sum_{i \epsilon I} P(x_i) * x_i \,, & if\ X\ is\ discrete \\ \int_{x \epsilon \mathbb{R}} p(x) * x\ * dx \,, & if\ X\ is\ continuous \end{cases}$$

Mean(X) := AverageValue(X) :=$\mathbb{E}[X]$

# Probability Review

Suppose that X is a random variable with probability distribution P(x) or probability density p(x). Then we define:

$$\text{Expected value} = \mathbb{E}[X] := \begin{cases} \sum_{i \epsilon I} P(x_i) * x_i \,, & if\ X\ is\ discrete \\ \int_{x \epsilon \mathbb{R}} p(x) * x\ * dx \,, & if\ X\ is\ continuous \end{cases}$$

Mean(X) := AverageValue(X) :=$\mathbb{E}[X]$

The expected value / mean is a linear function for sums and (scalar) multiples of random variables!

# Probability Review

Suppose that X is a random variable with probability distribution P(x) or probability density p(x). Then we define:

$$\text{Expected value} = \mathbb{E}[X] := \begin{cases} \sum_{i \epsilon I} P(x_i) * x_i \,, & if\ X\ is\ discrete \\ \int_{x \epsilon \mathbb{R}} p(x) * x\ * dx \,, & if\ X\ is\ continuous \end{cases}$$

Mean(X) := AverageValue(X) :=$\mathbb{E}[X]$

The expected value / mean is a linear function for sums and (scalar) multiples of random variables!

Variance := Var(X) :=

$$\mathbb{E}[\ (X\ -\ \mathbb{E}[X])^2\ ] = \begin{cases} \sum_{i \epsilon I} P(x_i) * (x_i\ - \mathbb{E}[X])^2 \,, & if\ X\ is\ discrete \\ \sum_{x \epsilon \mathbb{R}} p(x) * (x\ - \mathbb{E}[X])^2 * dx \,, & if\ X\ is\ continuous \end{cases}$$

# Probability Review

Suppose that X is a random variable with probability distribution P(x) or probability density p(x). Then we define:

$$\text{Expected value} = \mathbb{E}[X] := \begin{cases} \sum_{i \epsilon I} P(x_i) * x_i \,, & if\ X\ is\ discrete \\ \int_{x \epsilon \mathbb{R}} p(x) * x\ * dx \,, & if\ X\ is\ continuous \end{cases}$$

Mean(X) := AverageValue(X) :=$\mathbb{E}[X]$

The expected value / mean is a linear function for sums and (scalar) multiples of random variables!

Variance := Var(X) :=

$$\mathbb{E}[\ (X\ -\ \mathbb{E}[X])^2\ ] = \begin{cases} \sum_{i \epsilon I} P(x_i) * (x_i\ - \mathbb{E}[X])^2 \,, & if\ X\ is\ discrete \\ \sum_{x \epsilon \mathbb{R}} p(x) * (x\ - \mathbb{E}[X])^2 * dx \,, & if\ X\ is\ continuous \end{cases}$$

The variance can also be expressed as $\mathbb{E}[X^2] - \mathbb{E}[\ X\ ]^2$, so we see that Var(X) = Var(-X) and Var(X) >= 0.

# Probability Review

Suppose that X is a random variable with probability distribution P(x) or probability density p(x). Then we define:

$$\text{Expected value} = \mathbb{E}[X] := \begin{cases} \sum_{i \epsilon I} P(x_i) * x_i \,, & if\ X\ is\ discrete \\ \sum_{x \epsilon \mathbb{R}} p(x) * x\ * dx \,, & if\ X\ is\ continuous \end{cases}$$

Mean(X) := AverageValue(X) := $\mathbb{E}[X]$

The expected value/ mean is a linear function for sums and (scalar) multiples of random variables!

Variance := Var(X) :=

$$\mathbb{E}[\ (X\ -\ \mathbb{E}[X])^2\ ] = \begin{cases} \sum_{i \epsilon I} P(x_i) * (x_i\ - \mathbb{E}[X])^2 \,, & if\ X\ is\ discrete \\ \sum_{x \epsilon \mathbb{R}} p(x) * (x\ - \mathbb{E}[X])^2 * dx \,, & if\ X\ is\ continuous \end{cases}$$

The variance can also be expressed as $\mathbb{E}[X^2] - \mathbb{E}[\ X\ ]^2$, so we see that Var(X) = Var(-X) and Var(X) >= 0.

Standard Deviation := StdDev(X) := sqrt(Var(X))

# Probability Review

We also have the covariance of two random variables X and Y

$$\text{Cov(X,Y)} := \mathbb{E}[(X - \mathbb{E}[X]) * (Y - \mathbb{E}[Y])]$$

# Probability Review

We also have the covariance of two random variables X and Y

$$\text{Cov(X,Y)} := \mathbb{E}[(X - \mathbb{E}[X]) * (Y - \mathbb{E}[Y])]$$

Note Cov(X, Y) = E[XY] - E[X] * E[Y] is symmetric and linear in each of the variables. In particular, if c is a constant then

Cov(X, c) = 0

Cov(X + c , Y) = Cov(X,Y)

Cov(X + Z , Y) = Cov(X,Y) + Cov(Z,Y)

Cov(c*X , Y) = c * Cov(X,Y) = Cov(X, c*y).

# Probability Review

We also have the covariance of two random variables X and Y

$$\text{Cov(X,Y)} := \mathbb{E}[(X - \mathbb{E}[X]) * (Y - \mathbb{E}[Y])]$$

Note Cov(X, Y) = E[XY] - E[X] * E[Y] is symmetric and linear in each of the variables. In particular, if c is a constant then

Cov(X, c) = 0

Cov(X + c , Y) = Cov(X,Y)

Cov(X + Z , Y) = Cov(X,Y) + Cov(Z,Y)

Cov(c*X , Y) = c * Cov(X,Y) = Cov(X, c*y).

From the definition we see that Var(X) = Cov(X,X) is quadratic in X under scaling. In general for two random variables

Var(X + Y) = Var(X) + Var(Y) + 2*Cov(X,Y).

# Probability Review

We also have the covariance of two random variables X and Y

$$\text{Cov(X,Y)} := \mathbb{E}[(X - \mathbb{E}[X]) * (Y - \mathbb{E}[Y])]$$

Note Cov(X, Y) = E[XY] - E[X] * E[Y] is linear in each of the variables (i.e. if c is a constant then

Cov(c*X , Y) = c * Cov(X,Y) = Cov(X, c*y)

From the definition we see that Var(X) = Cov(X,X) is quadratic in X under scaling, and in general for two random variables Var(X + Y) = Var(X) + Var(Y) + 2*Cov(X,Y).

When X and Y are independent (i.e. the joint probability p(x,y) = p(x) * p(y) ) we have that Cov(X,Y) = 0 so Var(X+Y) = Var(X) + Var(Y).

The **correlation** between two random variables is closely related to the covariance between them. Since

$$|\text{Cov(X, Y)}| < \text{sqrt(Var(X) * Var(Y))}$$

# Probability Review

We also have the covariance of two random variables X and Y

$$\text{Cov(X,Y)} := \mathbb{E}[(X - \mathbb{E}[X]) * (Y - \mathbb{E}[Y])]$$

Note Cov(X, Y) = E[XY] - E[X] * E[Y] is linear in each of the variables (i.e. if c is a constant then

Cov(c*X , Y) = c * Cov(X,Y) = Cov(X, c*y)

From the definition we see that Var(X) = Cov(X,X) is quadratic in X under scaling, and in general for two random variables Var(X + Y) = Var(X) + Var(Y) + 2*Cov(X,Y).

When X and Y are independent (i.e. the joint probability p(x,y) = p(x) * p(y) ) we have that Cov(X,Y) = 0 so Var(X+Y) = Var(X) + Var(Y).

The **correlation** between two random variables is closely related to the covariance between them. Since

$$|\text{Cov(X, Y)}| < \text{sqrt(Var(X) * Var(Y))}$$

We see that the correlation

$$\text{Corr(X,Y)} := \text{Cov(X,Y)} / (\text{StdDev(X)} * \text{StdDev(Y)})$$

is always between -1 and 1.

# Probability Review

We often perform the operation of **translation** and **scaling** on random variable X to transform them into a normalized form:

# Probability Review

We often perform the operation of **translation** and **scaling** on random variable X to transform them into a normalized form:

**Shifting/Translating:** This adds a constant to the value of the random variable:

$$X \rightarrow X + c$$

This changes the mean $E[X + c] = E[X] + E[c] = E[x] + c$, but not the variance.

# Probability Review

We often perform the operation of **translation** and **scaling** on random variable X to transform them into a normalized form:

**Shifting/Translating:** This adds a constant to the value of the random variable:

$X \rightarrow X + c$

This changes the mean $E[X + c] = E[X] + E[c] = E[x] + c$, but not the variance.

**Scaling:** This multiplies the value of the random variable by a constant:

$X \rightarrow c*X$ for some (usually positive) non-zero constant c.

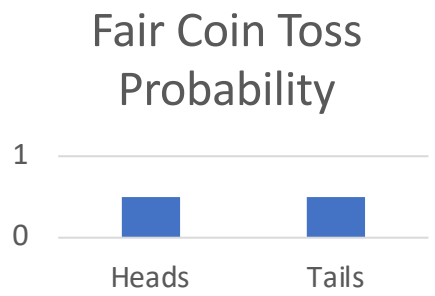
This changes both the mean $E[c*X] = c*E[X]$ and the variance $Var(c*X) = c^2*Var(X)$.

Note: To simplify our analysis or to ensure different variables are in comparable units, we often we normalize our random variables to have mean zero and variance 1, e.g. transforming any normal/Gaussian distribution into the standard normal distribution N(0,1).

# Common Discrete Probability Distributions / RVs

1) Bernoulli Distribution – (i.e. weighted coin toss)

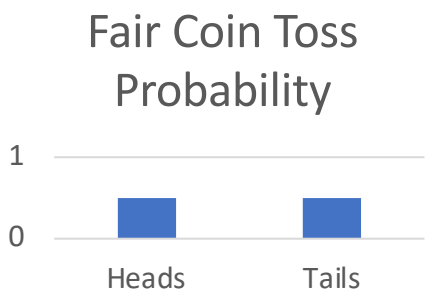
This takes two possible values, 0 and 1, and is described by a parameter/weight p with 0 <= p <= 1. Here P(1) = p and P(0) = 1-p.



# Common Discrete Probability Distributions / RVs

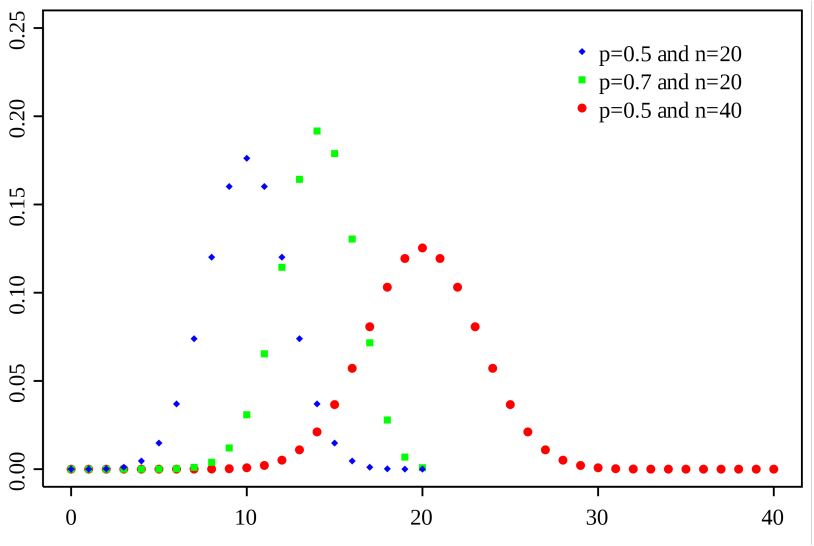
1) Bernoulli Distribution – (i.e. weighted coin toss)

This takes two possible values, 0 and 1, and is described by a parameter/weight p with 0 <= p <= 1. Here P(1) = p and P(0) = 1-p.




2) Binomial Distribution – (i.e. the sum of k weighted coin tosses)

This takes k+1 possible values, 0, 1, 2, 3, …, k, and is described by a parameter/weight p with 0 <= p <= 1 and a positive integer k.
The probability is given by $P(m) = p^m(1-p)^{k-m}\binom{k}{m}$ and $\binom{k}{m}$ is the binomial coefficient $\binom{k}{m} = \frac{k!}{m!(k-m)!}$ and $k! = k(k-1)(k-2)\cdots 2 \cdot 1$ denotes "k factorial".

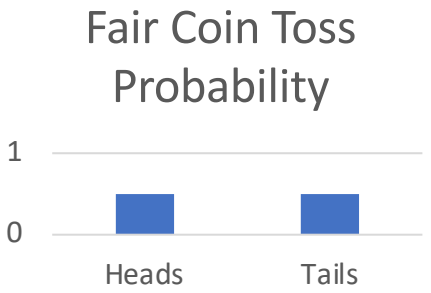



Images courtesy of Wikipedia.org

# Common Discrete Probability Distributions / RVs

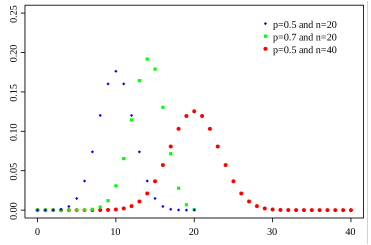
1) Bernoulli Distribution – (i.e. weighted coin toss)

This takes two possible values, 0 and 1, and is described by a parameter/weight p with 0 <= p <= 1. Here P(1) = p and P(0) = 1-p.



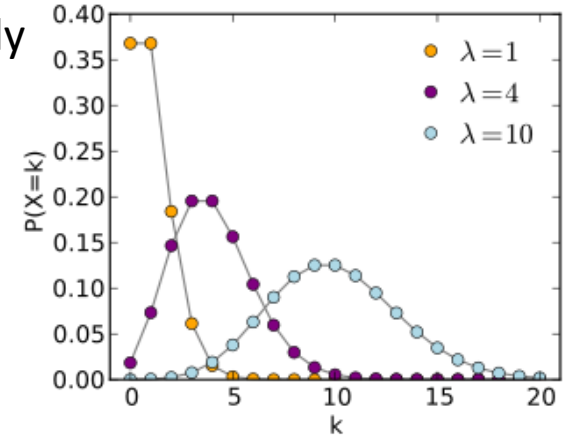

2) Binomial Distribution – (i.e. the sum of k weighted coin tosses)

This takes k+1 possible values, 0, 1, 2, 3, …, k, and is described by a parameter/weight p with 0 <= p <= 1 and a positive integer k. The probability is given by $P(m) = p^m(1-p)^{k-m}\binom{k}{m}$ and $\binom{k}{m}$ is the binomial coefficient $\binom{k}{m} = \frac{k!}{m!(k-m)!}$ and $k! = k(k-1)(k-2)\cdots 2 \cdot 1$ denotes "k factorial".




3) Poisson Distribution – the count of independent identically distributed events occurring in a given window of time.

This is described by a rate parameter r > 0 and has event probability given by $p(k) = \frac{r^k e^{-k}}{k!}$.

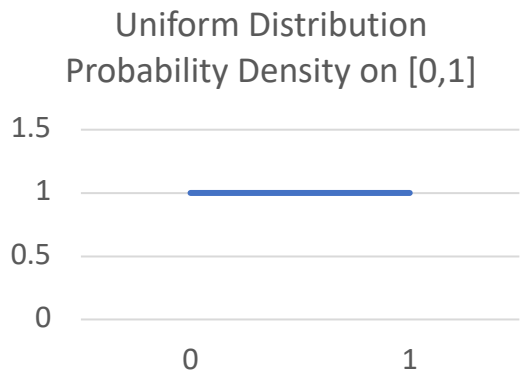



Images courtesy of Wikipedia.org

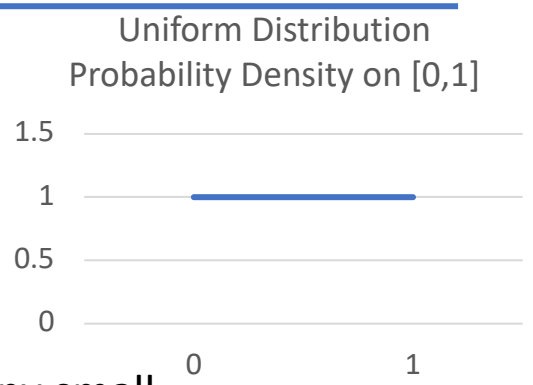
# Common Continuous Probability Distributions / RVs

4) Uniform Distribution – values distributed uniformly on [a,b]

No parameters, or two parameters a <= b.
The probability density function is given by p(x) = 1/(b-a).



# Common Continuous Probability Distributions / RVs



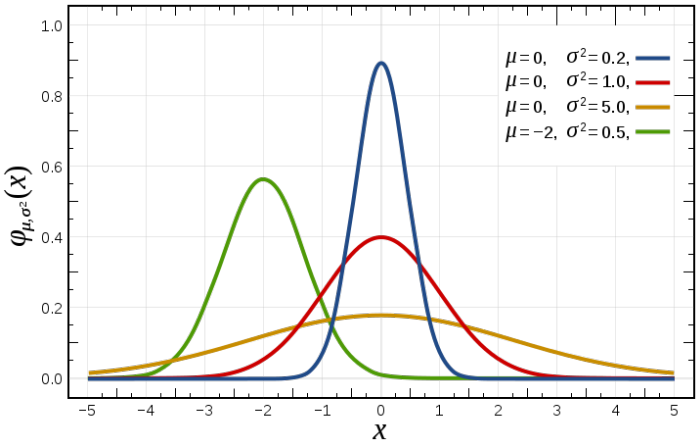

4) Uniform Distribution – values distributed uniformly on [a,b]

No parameters, or two parameters a <= b.
The probability density function is given by p(x) = 1/(b-a).

5) Normal Distribution – the cumulative effect (i.e. sum) of many small independent variables. (By the Central Limit Theorem)

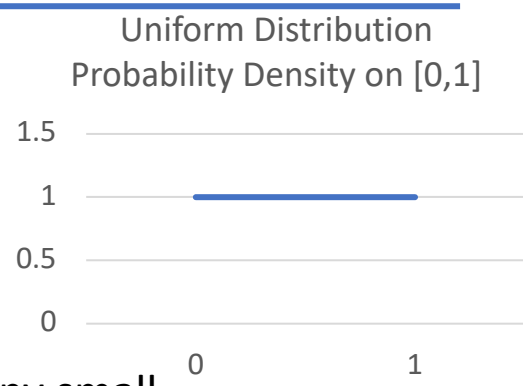
Described by the parameters $\mu$ (mean) and $\sigma^2$ (variance).
The probability density function is given by

$$p(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{\sigma^2}} .$$




Images courtesy of Wikipedia.org

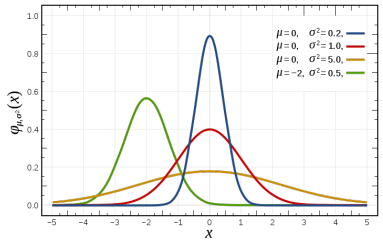
# Common Continuous Probability Distributions / RVs




4) Uniform Distribution – values distributed uniformly on [a,b]

No parameters, or two parameters a <= b.
The probability density function is given by p(x) = 1/(b-a).

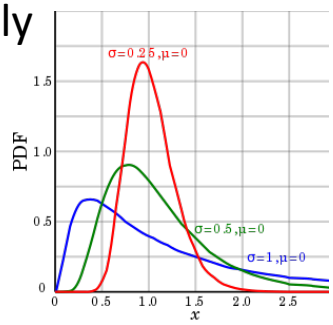
5) Normal Distribution – the cumulative effect (i.e. sum) of many small independent variables. (By the Central Limit Theorem)




Described by the parameters $\mu$ (mean) and $\sigma^2$ (variance).
The probability density function is given by

$$p(x) = \frac{1}{\sqrt{2\cdot\pi\cdot\sigma^2}}\,\mathrm{e}^{-\frac{(x-\mu)^2}{\sigma^2}}\ .$$

**6**) Log-normal Distribution – the distribution of exp(X) where X is normally distributed.



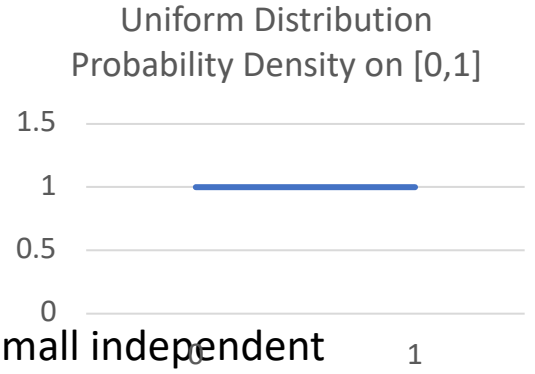

Described by the parameters mu (mean) and sigma-squared (variance).

Images courtesy of Wikipedia.org

# Common Continuous Probability Distributions / RVs

4) Uniform Distribution – values distributed uniformly on [a,b]

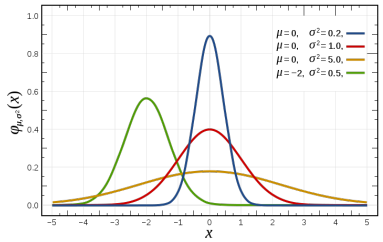
No parameters, or two parameters a <= b.
The probability density function is given by p(x) = 1/(b-a).




5) Normal Distribution – the cumulative effect (i.e. sum) of many small independent variables. (By the Central Limit Theorem)

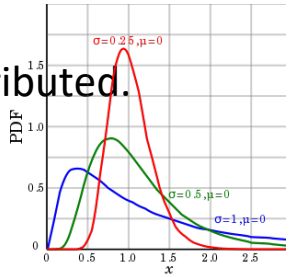
Described by the parameters $\mu$ (mean) and $\sigma^2$ (variance).
The probability density function is given by

$$p(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} e^{-\frac{(x-\mu)^2}{\sigma^2}} .$$

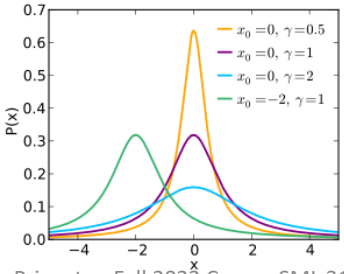



**6**) Log-normal Distribution – the distribution of exp(X) where X is normally distributed.

Described by the parameters mu (mean) and sigma-squared (variance).




7) Cauchy Distribution – described by two parameters, has fat tails, but no well-defined mean or variance.



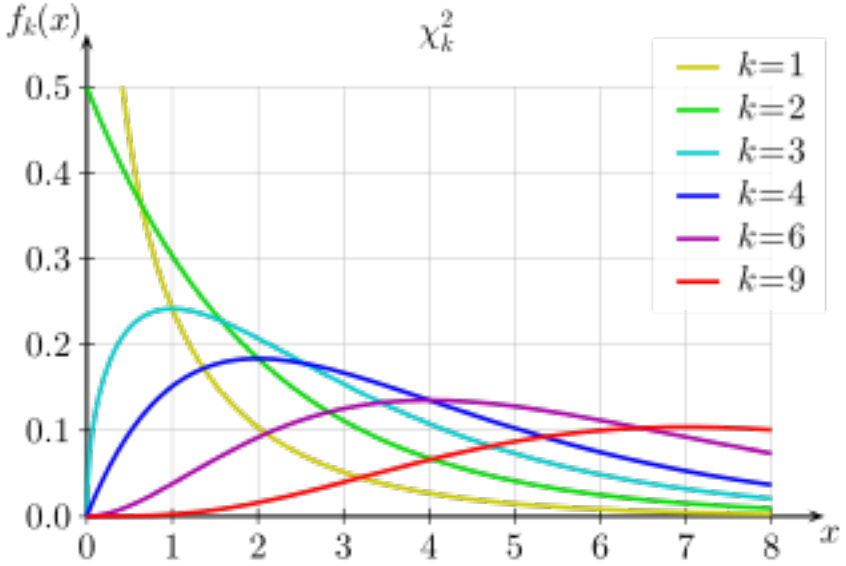

Images courtesy of Wikipedia.org

# Common Continuous Probability Distributions / RVs

8) Chi-squared Distribution – the distribution of the sum of squares of $k$ i.i.d. standard normal random variables $X_i \sim N(0,1)$.

Parameter given by a positive integer $k$.

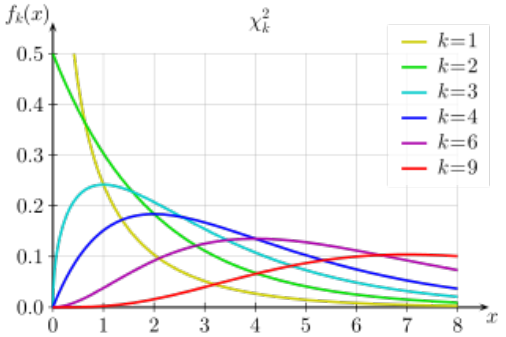
The probability density function is given by $p(x) = C\, x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$ for some $C > 0$.




Images courtesy of Wikipedia.org

# Common Continuous Probability Distributions / RVs

8) Chi-squared Distribution – the distribution of the sum of squares of $k$ i.i.d. standard normal random variables $X_i \sim N(0,1)$.




Parameter given by a positive integer $k$.

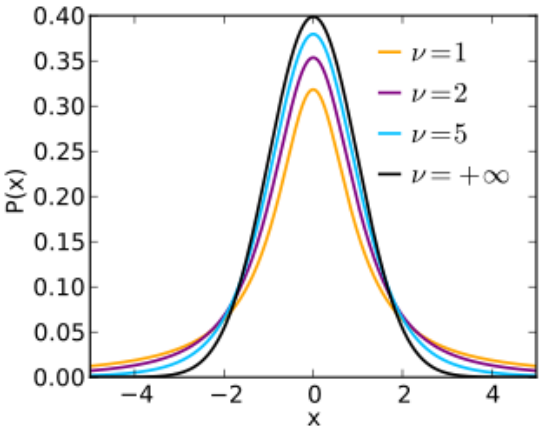
The probability density function is given by

$p(x) = C\, x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$ for some $C > 0$.

9) Student's T-Distribution – related to the normalized distribution of the true mean relative to the sample mean after $\nu + 1$ observations of the normal distribution.

Described by the parameter $\nu > 0$ (degrees of freedom). The probability density function is given by

$$p(x) = C \cdot \left(1 - \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$




Images courtesy of Wikipedia.org

# Statistics Review

**Statistics** is the mathematical study of events and their outcomes (weighted by probabilities). There can be finitely many or infinitely many possible outcomes. The mathematics of probability may look different in these two cases, but all ideas are the same.

Suppose we have a sample of n (i.i.d.) data points drawn from a random variable

# Statistics Review

**Statistics** is the mathematical study of events and their outcomes (weighted by probabilities). There can be finitely many or infinitely many possible outcomes. The mathematics of probability may look different in these two cases, but all ideas are the same.

Suppose we have a sample of n (i.i.d.) data points drawn from a random variable

1) Sample Mean – converges to the probability mean E[X] by the Law of Large Numbers.

# Statistics Review

**Statistics** is the mathematical study of events and their outcomes (weighted by probabilities). There can be finitely many or infinitely many possible outcomes. The mathematics of probability may look different in these two cases, but all ideas are the same.

Suppose we have a sample of n (i.i.d.) data points drawn from a random variable

1) Sample Mean – converges to the probability mean E[X] by the Law of Large Numbers.

2) Sample Variance / Covariance – estimator as you expect, but with 1/(N-1) to give more accuracy for smaller sample sizes.

# Statistics Review

**Statistics** is the mathematical study of events and their outcomes (weighted by probabilities). There can be finitely many or infinitely many possible outcomes. The mathematics of probability may look different in these two cases, but all ideas are the same.

Suppose we have a sample of n (i.i.d.) data points drawn from a random variable

1) Sample Mean – converges to the probability mean E[X] by the Law of Large Numbers.

2) Sample Variance / Covariance – estimator as you expect, but with 1/(N-1) to give more accuracy for smaller sample sizes.

3) Estimator – a statistic that gives a way of estimating a property of the probability distribution from a sample of data drawn from it. Usually denoted with a "hat".

# Statistics Review

**Statistics** is the mathematical study of events and their outcomes (weighted by probabilities). There can be finitely many or infinitely many possible outcomes. The mathematics of probability may look different in these two cases, but all ideas are the same.

Suppose we have a sample of n (i.i.d.) data points drawn from a random variable

1) Sample Mean – converges to the probability mean E[X] by the Law of Large Numbers.

2) Sample Variance / Covariance – estimator as you expect, but with 1/(N-1) to give more accuracy for smaller sample sizes.

3) Estimator – a statistic that gives a way of estimating a property of the probability distribution from a sample of data drawn from it. Usually denoted with a "hat".

4) Error and Bias – The difference between the estimator and the parameter being estimated, for a sample and for its expected value.

# Application to Data Science

**Data Science** is the discipline/art/study of extracting insights from data. Often we assume that we're given a sample of data drawn from a given (*usually unknown!*) probability distribution that we'd like to understand in some meaningful way.

- **Exploratory Data Analysis (EDA**) – what can we see about the structure of the data by looking at it directly to identify patterns and internal structure. Also this might involve cleaning or transforming the data in some way for further analysis.

- Prediction of a dependent variable based on other features – regression or classification, or some other kind of modelling.

- Finding Hidden Structure – clustering, feature selection, dimensionality reduction, etc.

Probability and Statistics create a common language for talking about what we're interested in describing, but often act as more of a guiding principle than a rigorous approach. When precise information in known about the underlying data distribution and simple models are used, then more rigorous conclusions can be derived.