Mathematical Foundations: Probability Theory

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2022)
http://cs182.sist.shanghaitech.edu.cn

App. A of I2ML

Motivation

Question

Given: We have 25 Male and 15 Female students. If a student is randomly picked from these 2 groups, which group will you guess the student is from?

2 classes: C_1 = Male, C_2 = Female

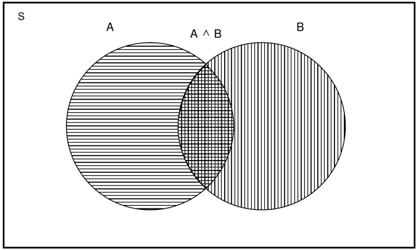
- the state of nature is unpredictable $\rightarrow$ use probability
- The set of all possible outcomes is known as the sample space S.
  - A sample space is discrete if it consists of a finite (or countably infinite) set of outcomes; otherwise it is continuous.
- Any subset E of S is an event. The probability of the event is denoted as $P(E)$.

Axioms for Probability

- All probabilities are between 0 and 1: $0 \leq P(A) \leq 1$
- The certain event has probability 1
- The impossible event has probability 0
- If A and B are any two events,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- S is the sample space containing all possible outcomes, $P(S) = 1$.



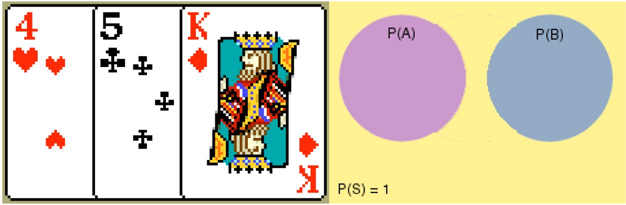
Mutually Exclusive Events

Two events are mutually exclusive if they cannot occur at the same time

Example

A single card is chosen at random from a standard deck of 52 playing cards

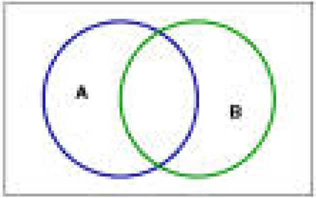
- A: the card chosen is a five, B: the card chosen is a king
- mutually exclusive? $A \cap B = \emptyset$.



$$P(A \cup B) = P(A) + P(B)$$

Conditional Probability - I

- Let A and B be two events such that $P(A) > 0$
- $P(B \mid A)$: probability of B given that A has occurred



$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \quad P(A \cap B) = P(B \mid A)P(A) \quad \text{(product rule)}$$

- probability that both A and B occur is equal to the probability that A occurs times the probability that B occurs given that A has occurred
- Because $\cap$ is commutative, we have $P(A \cap B) = P(A \mid B)P(B)$

Conditional Probability - II

- For any n events $A_1, A_2, \ldots, A_n$:

$$P(A_1 \cap A_2 \cap \cdots \cap A_{n-1} \cap A_n) = P(A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}) \cdots P(A_3 \mid A_1 \cap A_2) P(A_2 \mid A_1) P(A_1)$$

- (Formula of total probability or sum rule) If events $A_1, \ldots, A_n$ are mutually exclusive and exhaustive, i.e., $\bigcup_{i=1}^{n} A_i = S$ ($\sum_{i=1}^{n} P(A_i) = 1$), we have

$$B = \bigcup_{i=1}^{n} B \cap A_i$$

and then

$$\begin{aligned} P(B) =& P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n) \\ =& P(B \mid A_1) P(A_1) + P(B \mid A_2) P(A_2) + \cdots + P(B \mid A_n) P(A_n) \end{aligned}$$

Independence

Two events A and B are independent if

$$P(B \mid A) = P(B), \text{ or } P(A \mid B) = P(A)$$

Example

A and B are two coin tosses

- the probability of B occurring is not affected by the occurrence or non-occurrence of A
- knowledge about X contains no information about Y
- this is also equivalent to $P(A \cap B) = P(A)P(B)$

If n events $(A_1, \ldots, A_n)$ are independent

$$P(A_1 \cap \cdots \cap A_n) = \prod_{i=1}^{n} P(A_i)$$

Bayes' Theorem or Rule

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

or, generally,

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

$$P(C_i \mid x) = \frac{P(x \mid C_i)P(C_i)}{P(x)}$$

- $P(C_i)$: prior probability of C_i
  - initial probability for C_i, before observing the training data
- $P(C_i \mid x)$: posterior probability for C_i after observing the data x
- $P(x \mid C_i)$: likelihood of observing the data x given class C_i
- $P(x)$: probability that training data x will be observed

Example: Medical Diagnosis

Given:

- $P(\text{Cough} \mid \text{SARS}) = 0.8$
- $P(\text{SARS}) = 0.005$
- $P(\text{Cough}) = 0.05$

Question

Find: $P(\text{SARS} \mid \text{Cough})$

$$
\begin{aligned}
& P(\text{SARS} \mid \text{Cough}) \\
= & \frac{P(\text{Cough} \mid \text{SARS}) P(\text{SARS})}{P(\text{Cough})} \\
= & \frac{0.8 \times 0.005}{0.05} = 0.08
\end{aligned}
$$

Random Variables

- A random variable (RV) is a function that assigns a number to each outcome in the sample space S of a random experiment.
- The distribution function $F(\cdot)$ of a random variable X for any real number x is

$$F(x) = P(X \leq x)$$

and we have

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$

Discrete Probability Distributions

X: discrete random variable

Probability mass function (pmf) (or probability function, probability distribution)

$$p(x) = P(X = x)$$

Cumulative distribution function (or distribution function):

$$F(x) = P(X \leq x)$$

- if X takes on only a finite number of values $x_1, x_2, \ldots x_n$

$$F(x) = \begin{cases} 0 & -\infty < x < x_1 \\ P(X = x_1) & x_1 \leq x < x_2 \\ P(X = x_1) + P(X = x_2) & x_2 \leq x < x_3 \\ \vdots & \vdots \\ P(X = x_1) + \cdots + P(X = x_n) & x_n \leq x < \infty \end{cases}$$

Continuous Probability Distributions

X: continuous random variable

- $p(x)$: probability density function (pdf) (or probability function, probability distribution)
- the probability that X lies between two different values is more meaningful

$$P(a < X < b) = \int_a^b p(x)dx$$

  - the probability that X takes on any one particular value is generally zero
- Cumulative distribution function:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(x)dx$$

  and

$$\frac{dF(x)}{dx} = p(x)$$

Joint Distributions: Discrete - I

- generalization to two or more random variables
- if X and Y are two discrete random variables, we define the joint probability mass function of X and Y by

$$P(X = x, Y = y) = p(x, y)$$

  where $p(x, y) \geq 0$ and $\sum_x \sum_y p(x, y) = 1$
- marginal probability mass function of X

$$p_X(x) = P(X = x) = \sum_j p(x, y_j)$$

Joint Distributions: Discrete - II

- joint distribution function

$$F(x,y) = P(X \le x, Y \le y) = \sum_{u \le x} \sum_{v \le y} p(u,v)$$

- marginal distribution function of X

$$F_X(x) = P(X \le x) = \sum_{u \le x} \sum_{j} p(u, v_j)$$

and

$$F_X(x) = F(x, \infty)$$

Joint Distributions: Continuous - I

- if X and Y are continuous random variables, the joint density function of X and Y is

$$p(x, y)$$

where $p(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) dxdy = 1$ and

$$P(a < X < b, c < Y < d) = \int_{x=a}^{b} \int_{y=c}^{d} p(x, y) dxdy$$

- marginal probability density function of X

$$p_X(x) = \int_{v=-\infty}^{\infty} p(x, v) dv$$

Joint Distributions: Continuous - II

- joint distribution function

$$F(x,y) = P(X \leq x, Y \leq y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} p(u,v)dudv$$

$$\frac{\partial^2 F}{\partial x \partial y} = p(x,y)$$

- marginal distribution function of X

$$F_X(x) = P(X \leq x) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{\infty} p(u,v)dudv$$

and

$$F_X(x) = F(x,\infty)$$

Conditional Distributions and Bayes' Theorem

- if X and Y are random variables,

$$p(x \mid y) = p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x, y)}{p_Y(y)}$$

- if X and Y are independent, we have

$$p(x, y) = p_X(x) p_Y(y)$$

  and

$$F(x, y) = F_X(x) F_Y(y)$$

- When X and Y are jointly distributed with the value of X known, the probability that Y takes a given value can be computed using Bayes' rule

$$p(y \mid x) = \frac{p(x \mid y) p_Y(y)}{p_X(x)} = \frac{p(x \mid y) p_Y(y)}{\sum_y p(x \mid y) p_Y(y)}$$

Mathematical Expectation

- aka expected value or expectation or mean of a random variable X
  - X discrete:

$$\mathrm{E}(X) = \sum_{j=1}^{n} x_j P(X = x_j)$$

  - X continuous:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x p(x) dx$$

- Properties:

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$$

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$

- For any real-valued function $g(\cdot)$
  - X discrete: $\mathrm{E}[g(X)] = \sum_{j=1}^{n} g(x_j) P(X = x_j)$
  - X continuous: $\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) p(x) dx$

Moments

rth moment: $\mathrm{E}(X^r)$

- mean $\mu = \mathrm{E}(X)$: 1st moment

rth central moment: $m_r = \mathrm{E}[(X-\mu)^r]$

- $m_0 = 1$, $m_1 = 0$
  - $m_2 = \mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \sigma^2$ is the variance; σ is the standard deviation (has the same unit as X)

Property of variance:

- $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$

Moments

covariance indicates the relationship between two random variables

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] = \mathrm{E}(XY) - \mu_X \mu_Y$$

Property of covariance:

- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X \pm Z, Y) = \mathrm{Cov}(X, Y) \pm \mathrm{Cov}(Z, Y)$
- $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X, Y)$
- $\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i,j=1}^{N} \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{N} \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$

If X and Y are independent, $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y) = \mu_X \mu_Y$ and $\mathrm{Cov}(X, Y) = 0$

correlation is a normalized, dimensionless quantity that is always between -1 and 1:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Moments

For multivariate random vector $\mathbf{X} = [X_1, X_2, \ldots, X_n]^T$:

- 2nd central moment: covariance matrix

$$\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{X}) = \mathrm{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]$$

where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$.

Covariance Matrix Example

For a 2-D vector $\mathbf{X} = [X_1, X_2]^T$:

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathrm{E}\left(\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix}^T \right) \\
&= \mathrm{E}\left(\begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 \end{bmatrix} \right) \\
&= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}
\end{aligned}$$

Weak Law of Large Numbers

Let $\{X^i\}_{i=1}^N$ be N independent and identically distributed (iid) random variables each having mean μ and a finite variance σ^2. Then for any $\epsilon > 0$,

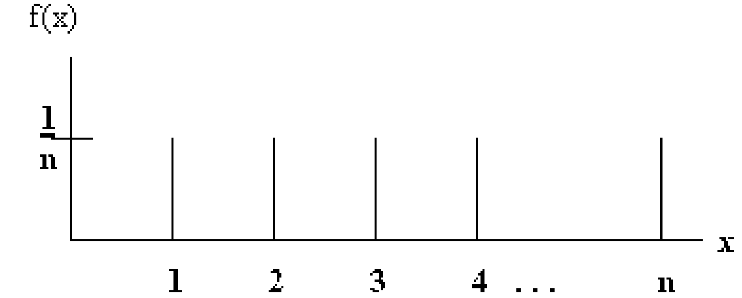
$$P\left(\left|\frac{\sum_{i=1}^N X^i}{N} - \mu\right| > \epsilon\right) \to 0 \quad \text{as } N \to \infty$$

Discrete RV Distribution Example: Uniform Distribution

Example

outcome of throwing a fair die

- $P(X = 1) = P(X = 2) = \cdots = P(X = 6)$



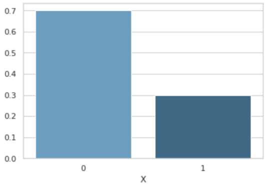
Discrete RV Distribution Example: Bernoulli Distribution

The Bernoulli random variable X is a $0/1$ indicator variable and takes the value 1 for a success outcome and is 0 otherwise. p is the probability that the result of trial is a success. Then

$$P(X = 1) = p \text{ and } P(X = 0) = 1 - p$$

or, equivalently,

$$P(X = x) = Ber(x; p) = p^x(1 - p)^{1-x}, \; x = 0, 1$$

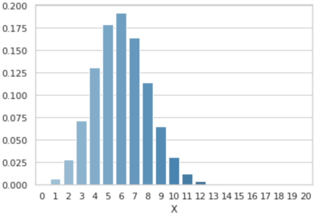


$$\mathrm{E}(X) = p \text{ and } \mathrm{Var}(X) = p(1 - p)$$

Discrete RV Distribution Example: Binomial Distribution

given: probability of getting a head is p, #heads when the biased coin is tossed N times (i.e, N iid Bernoulli trials)

$$P(X = x) = Bin(x; N, p) = \binom{N}{x} p^x (1-p)^{N-x} \text{ with } \binom{N}{x} = \frac{N!}{x!(N-x)!}$$



the distribution gets a nice bell shape

$$\mathrm{E}(X) = Np \text{ and } \mathrm{Var}(X) = Np(1-p)$$

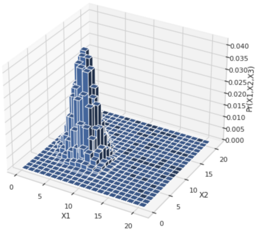
Discrete RV Distribution Example: Multinomial Distribution

given: probability of getting number i for $i = 1, \ldots, K$ is p_i with $\sum_{i=1}^{K} p_i = 1$ from rolling a die, # of i when the biased dice is rolled N times (i.e, N iid generalized Bernoulli trials)

$$P(X_1 = x_1, \ldots, X_K = x_K) = Mul(x_1, \ldots, x_K; N, p_1, \ldots, p_K) = N! \prod_{i=1}^{K} \frac{p_i^{x_i}}{x_i!}$$

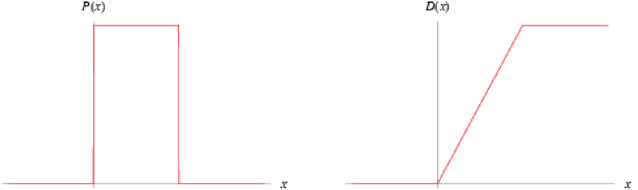
where outcome i occurred N_i times with $\sum_{i=1}^{K} x_i = N$. When $N = 1$,

$$P(x_1, \ldots, x_K; 1, p_1, \ldots, p_K) = \prod_{i=1}^{K} p_i^{x_i}$$



Continuous RV Distribution Example: Uniform Distribution

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x \leq b \\ 0 & \text{otherwise} \end{cases}$$



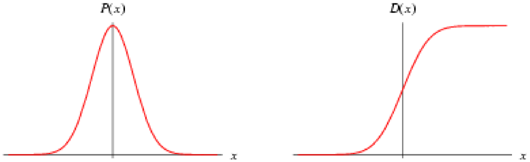
$$\mathrm{E}(X) = \frac{a+b}{2} \text{ and } \mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

Continuous RV Distribution Example: Normal (Gaussian) Distribution

for a RV X follows, i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$, its density function is

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

where μ is the location parameter and σ is the scale parameter.
For normal RVs, uncorrelatedness implies independence.



the distribution gets a nice bell shape

$$\mathrm{E}(X) = \mu \text{ and } \mathrm{Var}(X) = \sigma^2$$

z-normalization: a standard/unit normal RV $z = \frac{x-\mu}{\sigma} \sim \mathcal{N}(0,1) = \mathcal{Z}$

Central Limit Theorem

Let $X_1, X_2, \ldots, X_N$ be a set of iid RVs all having mean μ and variance σ^2. Then for large N, the distribution of

$$X_1 + X_2 + \cdots + X_N$$

is approximately $\mathcal{N}(N\mu, N\sigma^2)$.

Continuous RV Distribution Example: Chi-Square Distribution

If Z_i are independent unit normal RVs, then

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

follows a standard chi-square distribution with n degrees of freedom, namely, $X \sim \chi_n^2$, with

$$\mathrm{E}(X) = n \text{ and } \mathrm{Var}(X) = 2n$$

Continuous RV Distribution Example: t Distribution - I

If $Z \sim \mathcal{Z}$ and $X \sim \chi^2_n$ are independent RVs, then

$$T = \frac{Z}{\sqrt{X/n}}$$

follows a standard t distribution (or Student's t distribution) with n degrees of freedom, denoted as $T \sim \mathcal{T}_n$, with

$$\mathrm{E}(T) = 0,\ n > 1 \text{ and } \mathrm{Var}(T) = \frac{n}{n-1},\ n > 2$$

Like the standard normal density, t density is symmetric around 0. As n becomes larger, t density becomes more and more like the standard normal, the difference being that t has thicker tails, indicating greater variability than does normal.

Continuous RV Distribution Example: t Distribution - II

A standard t distribution is given by

$$p(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

where ν is the number of degrees of freedom and $\Gamma(\cdot)$ is the gamma function.
A general t distribution is given by

$$p(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}\sigma} \left(1 + \frac{1}{n}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{n+1}{2}}$$

where μ is the location parameter and σ is the scale parameter. We also have

$$\mathrm{E}(X) = \mu,\ n > 1 \text{ and } \mathrm{Var}(X) = \frac{n}{n-2}\sigma^2,\ n > 2$$

Continuous RV Distribution Example - I

Let us say $\{X^t\}_{t=1}^N$ are iid and follow $\mathcal{N}(\mu, \sigma^2)$. The estimated sample mean is

$$m = \frac{\sum_{t=1}^N X^t}{N}$$

and we have $\frac{m-\mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$. The estimated sample variance is

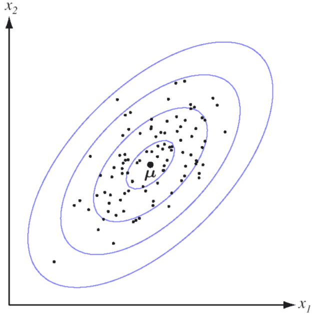
$$S^2 = \frac{\sum_{t=1}^N (X^t - m)^2}{N-1}$$

and we have $\frac{S^2}{\sigma^2/(N-1)} \sim \chi^2_{N-1}$.
It can be proved that m and S^2 are independent. Then, we can obtain

$$\frac{m-\mu}{S/\sqrt{N}} \sim \mathcal{T}_{N-1}$$

Multivariate RV Distribution Example: Normal (Gaussian) Distribution

- Random vector: $\mathbf{X} = [X_1, X_2, \ldots, X_n]^T$
- multivariate Gaussian: $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$



$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

We also have

$$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu} \text{ and } \mathrm{Var}(X) = \boldsymbol{\Sigma}$$

Multivariate RV Distribution Example: t Distribution

A general multivariate t distribution with p-variate is given by

$$p(\mathbf{x}) = \frac{\Gamma(\frac{n+p}{2})}{\Gamma(\frac{n}{2})(n\pi)^{\frac{p}{2}}\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}}\left(1 + \frac{1}{n}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-\frac{n+p}{2}}$$

where n is the degree of freedom (a shape parameter), $\boldsymbol{\mu}$ is the location parameter, and $\boldsymbol{\Sigma}$ is the scale parameter. We denote it as $\mathbf{x} \sim \mathcal{T}(n, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. We also have

$$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu},\ n > 1 \text{ and } \mathrm{Var}(X) = \frac{n}{n-2}\boldsymbol{\Sigma},\ n > 2$$