Discrete Probability Theory
revision course

Jonas Hübotter

Outline

Counting

Probability

Conditional probability

Discrete random variables

Continuous random variables

Inductive Statistics

Markov chains

Plan I

Counting

- Sample spaces and events
- Counting sets

Sample spaces and events

Definition 1

A sample space is the set of all possible outcomes of an experiment.

Definition 2

An event is a subset of the sample space.

Naive definition of probability of an event A in sample space S:

$$P(A) = \frac{\#\text{ favorable outcomes}}{\#\text{ possible outcomes}} = \frac{|A|}{|S|}$$

Assumptions:

- all outcomes equally likely
- finite sample space

Counting sets

Multiplication rule

Consider $i \in [m]$ experiments with n_i possible outcomes. Then the overall number of possible outcomes is

$$\prod_{i=1}^{m} n_i .$$

Sampling table

Given n objects, select k objects.

	order	$\neg$ order
replacement	n^k	$\binom{n+k-1}{k}$
$\neg$ replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Plan I

Probability

- σ-algebras
- Probability spaces
- Joint and marginal probabilities

σ-algebras

Definition 3

Given the set S. The set $\mathcal{A} \subseteq \mathcal{P}(S)$ is a σ-algebra over S if the following properties are satisfied:

- $S \in \mathcal{A}$;
- if $A \in \mathcal{A}$, then $\bar{A} \in \mathcal{A}$; and
- $\forall n \in \mathbb{N}.\ A_n \in \mathcal{A} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

Why do we need σ-algebras?

To describe events in the context of a probability space.

Probability spaces

Definition 4

Given the set S and the σ-algebra $\mathcal{A}$ over S. The function

$$P : \mathcal{A} \to [0, 1]$$

is a probability measure on $\mathcal{A}$ if the Kolmogorov axioms are satisfied:

- $P(S) = 1$;
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ if $\forall i \neq j.\ A_i \cap A_j = \emptyset$.

Definition 5

For an event $A \in \mathcal{A}$, $P(A)$ is the probability of A.

Definition 6

A probability space consists of

- a sample space S;
- a σ-algebra $\mathcal{A}$ over S; and
- a probability measure P on $\mathcal{A}$.

For a probability space the following properties hold:

- $P(\emptyset) = 0$
- $P(S) = 1$
- $0 \leq P(A) \leq 1$ for all $A \in \mathcal{A}$
- $P(\bar{A}) = 1 - P(A)$ for all $A \in \mathcal{A}$
- if $A, B \in \mathcal{A}$ and $A \subseteq B$, then $P(A) \leq P(B)$

Also the principle of inclusion-exclusion holds:

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{I \subseteq [n], I \neq \emptyset} (-1)^{|I|+1} \cdot P(\bigcap_{i \in I} A_i).$$

And Boole's inequality holds:

$$P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i).$$

Joint and marginal probabilities

A marginal probability is the probability of a single event irrespective of other events.

A joint probability is the probability of two or more events occurring simultaneously:

$$P(A, B) = P(A \cap B).$$

Plan I

Conditional probability

- Prior and posterior
- Independence
- Conditioning

Prior and posterior

Conditional probability *updates* the probability of an event A given some new information B.

$P(A)$ is called the prior and $P(A|B)$ the posterior probability.

$$P(A|B) = \frac{P(A, B)}{P(B)}.$$

The posterior is the joint probability of the event A and the information B relative to the probability of the information B.

Independence

Two events are independent if the occurrence of one event does not affect the probability of occurrence of the other event.

Two events A and B are independent
$\iff P(A|B) = P(A)$ for $P(B) > 0$
$\iff P(B|A) = P(B)$ for $P(A) > 0$
$\iff P(A, B) = P(A)P(B)$.

Conditioning

Some properties immediately follow from the definition of conditional probability:

- $P(A, B) = P(B)P(A|B) = P(A)P(B|A)$ as $A \cap B = B \cap A$
- $P(A_1, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \ldots, A_{n-1})$ (multiplication rule)
- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ (Bayes' rule)
- $P(A) = P(A, B) + P(A, \bar{B}) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$ (law of total probability)

Plan I

Discrete random variables

- Cumulative Distribution Function
- Probability Mass Function
- Independence
- Bernoulli
- Averages
- Indicator variables
- Binomial
- Variance
- Geometric
- Poisson
- Probability-generating functions
- Moment-generating functions
- Joint distributions
- Conditional distributions

Plan II

Convolutions
More distributions
Inequalities

Discrete random variables

Definition 7

A random variable X is a function

$$X : S \to \mathbb{R}.$$

A random variable is discrete if its domain S is finite or countable infinite.

The range of a discrete random variable

$$X(S) = \{x \in \mathbb{R}.\ \exists A \in S.\ X(A) = x\}$$

is also discrete.

Cumulative Distribution Function

$X \leq x$ is an event.

Definition 8

The cumulative distribution function of a random variable X is defined as $F_X(x) = P(X \leq x) \in [0, 1]$.

Properties of CDFs:

- monotonically increasing
- right-continuous
- $F_X(x) \xrightarrow{x \to -\infty} 0$
- $F_X(x) \xrightarrow{x \to \infty} 1$

Therefore, $P(a < X \leq b) = F_X(b) - F_X(a)$.

Probability Mass Function

Definition 9

The probability mass function of a discrete random variable X is defined as $f_X(x) = P(X = x) \in [0, 1]$ where

$$\sum_{x \in X(S)} f_X(x) = 1.$$

The CDF of X can be obtained from the PDF of X by summing over the PDF

$$F_X(x) = \sum_{x' \leq x} f_X(x').$$

The PMF of X can be obtained from the CDF of X by identifying the *jumps* in the CDF

$$f_X(x) = F_X(x) - F_X(prev(x)).$$

Independence

Two random variables are independent if knowledge about the value of one random variable does not affect the probability distribution of the other random variable.

Two discrete random variables X and Y are independent
$\iff$ the events $X = x$ and $Y = y$ are independent
$\iff$ the events $X \leq x$ and $Y \leq y$ are independent.

Bernoulli

Definition 10 ($X \sim Bern(p)$)

A discrete random variable X is Bernoulli distributed with parameter p when $X(S) = \{0, 1\}$ and $P(X = 1) = p$.

Overview

- $E(X) = p$
- $Var(X) = p(1 - p)$
- $G_X(s) = 1 - p + ps$
- $M_X(s) = 1 - p + pe^s$

Averages

Definition 11

The expected value $E(X)$ of a random variable X is the arithmetic mean of a large number of a realizations of X.

$$\begin{aligned} E(X) &= \sum_{x \in X(S)} x \cdot P(X = x) \\ &= \sum_{A \in S} X(A) \cdot P(A). \end{aligned}$$

For infinite probability spaces absolute convergence of $E(X)$ is necessary for the existence of $E(X)$.

Properties of expected values:

- if $\forall A \in S.\ X(A) \leq Y(A)$, then $E(X) \leq E(Y)$ (monotonicity)
- $E(a \cdot X + b) = a \cdot E(X) + b$, $E(X + Y) = E(X) + E(Y)$ (linearity)
- $E(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} E(X_i)$ if $X_1, \ldots, X_n$ independent (multiplicativity).

Definition 12

$E(X^i)$ is called the i-th moment of the random variable X and $E((X - E(X))^i)$ is called the i-th central moment of X.

The law of the unconscious statistician (LOTUS) can be used to find the expected value of transformed random variables.

$$E(g(X)) = \sum_{x \in X(S)} g(x) \cdot P(X = x).$$

Indicator variables

Definition 13

Given an event A, the random variable $I_A \sim Bern(P(A))$ is the indicator variable of the event A.

Properties of indicator variables:

- $E(I_A) = P(A)$ (fundamental bridge)
- $E(I_{A_1} \cdots I_{A_n}) = P(A_1 \cap \cdots \cap A_n)$.

Binomial

Definition 14 ($X \sim Bin(n, p)$)

A discrete random variable X has the binomial distribution with parameters n and p when X models the #successes in n independent $Bern(p)$ trials.

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Overview

- $E(X) = np$
- $Var(X) = np(1-p)$
- $G_X(s) = (1 - p + ps)^n$
- $M_X(s) = (1 - p + pe^s)^n$

Variance

Definition 15

The variance $Var(X)$ of a random variable X is a measure of the absolute deviation of a random variable from its mean.

$$\begin{aligned} Var(X) &= E((X - E(X))^2) \\ &= E(X^2) - E(X)^2. \end{aligned}$$

$SD(X) = \sqrt{Var(X)}$ is the standard deviation of X.

Properties of variances:

- $Var(a \cdot X + b) = a^2 \cdot Var(X)$
- $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$ if $X_1, \ldots, X_n$ independent.

Geometric

Definition 16 ($X \sim Geom(p)$)

A discrete random variable X has the geometric distribution with parameter p when X models the #trials leading up to a success in independent $Bern(p)$ trials.

$$f_X(k) = p(1-p)^{k-1}, k \in \mathbb{N}. \qquad F_X(k) = 1 - (1-p)^{\lfloor k \rfloor}.$$

Overview

- $E(X) = \frac{1}{p}$
- $Var(X) = \frac{1-p}{p^2}$
- $G_X(s) = \frac{ps}{1-(1-p)s}$

Memorylessness

Completing x trials that are all failures does not change the probability of the next y trials to include a success.

This property can be formalized as follows:

$$P(X > y + x | X > x) = P(X > y).$$

The geometric distribution is the only memoryless discrete distribution.

Poisson

Definition 17 ($X \sim Po(\lambda)$)

A discrete random variable X has the Poisson distribution with parameter λ when X models the #events in a fixed interval with rate λ and with events independently occurring of the time since the last event.

$$f_X(k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}, k \in \mathbb{N}_0. \qquad F_X(k) = e^{-\lambda} \cdot \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}.$$

Overview

- $E(X) = \lambda$
- $Var(X) = \lambda$
- $G_X(s) = exp(\lambda(s-1))$
- $M_X(s) = exp(\lambda(e^s - 1))$

Poisson approximation to the Binomial

Let $X \sim Bin(n, \lambda/n)$.
Then the distribution of X converges to $Po(\lambda)$ as $n \to \infty$
(i.e. for small λ/n).

Probability-generating functions

Definition 18

Given a discrete random variable X with $X(S) \subseteq \mathbb{N}_0$ the probability-generating function is defined as

$$\begin{aligned} G_X(s) &= \sum_{x \in X(S)} s^x \cdot P(X = x) \\ &= E(s^X). \end{aligned}$$

The PGF of a random variable X generates the PMF of X:

$$P(X = i) = \frac{G_X^{(i)}(0)}{i!}.$$

Properties of probability-generating functions:

- $E(X) = G'_X(1)$
- $Var(X) = G''_X(1) + G'_X(1) - (G'_X(1))^2$
- $G_{X+t}(s) = s^t \cdot G_X(s), t \in \mathbb{N}_0$
- $G_{X+Y}(s) = G_X(s) \cdot G_Y(s)$ if X, Y independent
- $G_Z(s) = G_N(G_X(s))$ for $Z = X_1 + \cdots + X_N$, X_i i.i.d. with PGF G_X, and N independent.

Moment-generating functions

Definition 19

Given a random variable X the moment-generating function is defined as

$$\begin{aligned} M_X(s) &= \sum_{x \in X(S)} e^{sx} \cdot P(X = x) \\ &= E(e^{sX}) \\ &= \sum_{i=0}^{\infty} \frac{E(X^i)}{i!} \cdot s^i. \end{aligned}$$

The MGF of a random variable X generates the i-th moment of X:

$$E(X^i) = M_X^{(i)}(0).$$

Properties of moment-generating functions:

- $M_X(s) = G_X(e^s)$ if $X(S) \subseteq \mathbb{N}_0$
- $M_{X+Y}(s) = M_X(s) \cdot M_Y(s)$ if X, Y independent.

Joint distributions

Definition 20

A joint distribution is the distribution of two or more random variables.

$$f_{X,Y}(x,y) = P(X = x, Y = y).$$

The marginal distribution of a random variable can be obtained from a joint distribution by summing over all other random variables:

$$f_X(x) = \sum_{y \in Y(S)} f_{X,Y}(x,y).$$

Conditional distributions

Definition 21

Given the joint distribution of two random variables X and Y the conditional distribution of X given Y is the distribution of X when Y is known to be a particular value.

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{Y|X}(y|x) \cdot f_X(x)}{f_Y(y)}.$$

The conditional expectation of the random variables $X|Y = y$ is the expected value of the distribution $f_{X|Y=y}$:

$$E(X|Y = y) = \sum_{x \in X(S)} x \cdot f_{X|Y}(x|y).$$

Convolutions

Definition 22

Let X and Y be independent and $Z = X + Y$. Then

$$f_Z(z) = \sum_{x \in X(S)} f_X(x) \cdot f_Y(z - x).$$

The derivation of the distribution of a sum of random variables given the marginal distributions is called convolution.

More distributions

Definition 23 ($X \sim HypGeom(r, a, b)$)

A discrete random variable X has the hypergeometric distribution with parameters r, a and b when X models the # of drawn objects that have a specified feature in r draws without replacement from $a + b$ objects where b objects have the specified feature.

$$f_X(x) = \frac{\binom{b}{x}\binom{a}{r-x}}{\binom{a+b}{r}}.$$

Overview

- $E(X) = r \cdot \frac{b}{a+b}$

Definition 24 ($Z \sim NegBin(n, p)$)

A discrete random variable Z has the negative binomial distribution with parameters n and p when Z models the # of independent $Bern(p)$ trials before the n-th success.

$$f_Z(z) = \binom{z-1}{n-1} p^n (1-p)^{z-n}.$$

Example 25

Let $X_1, \ldots, X_n \sim Geom(p)$ i.i.d.
Then $Z = X_1 + \cdots + X_n \sim NegBin(n, p)$.

Inequalities

Inequalities vs approximations

Approximations allow us to model more complex problems but you usually don't know how good the approximation is.
Inequalities allow us to prove definite facts (i.e. bounds) about probabilities of certain events.

Definition 26 (Markov)

Given a random variable $X \geq 0$ and $t > 0$

$$P(X \geq t) \leq \frac{E(X)}{t}.$$

Definition 27 (Chebyshev)

Given a random variable X and $t > 0$

$$P(|X - E(X)| \geq t) \leq \frac{Var(X)}{t^2}.$$

Definition 28 (Chernoff)

Let $X_1, \dots, X_n$ be independent, Bernoulli-distributed random variables with $X_i \sim Bern(p_i)$. Then the following inequalities hold for $X = \sum_{i=1}^n X_i$ and $\mu = E(X) = \sum_{i=1}^n p_i$.

- $P(X \geq (1+\delta)\mu) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$ for all $\delta > 0$;
- $P(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^\mu$ for all $0 < \delta < 1$;
- $P(X \geq (1+\delta)\mu) \leq e^{-\mu\delta^2/3}$ for all $0 < \delta \leq 1$;
- $P(X \leq (1-\delta)\mu) \leq e^{-\mu\delta^2/2}$ for all $0 < \delta \leq 1$;
- $P(|X-\mu| \geq \delta\mu) \leq 2e^{-\mu\delta^2/3}$ for all $0 < \delta \leq 1$;
- $P(X \geq (1+\delta)\mu) \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu}$; and
- $P(X \geq t) \leq 2^{-t}$ for all $t \geq 2e\mu$.

Plan I

- Continuous random variables
  - Measure Theory
  - Continuous probability spaces
  - Uniform
  - Normal (Gaussian)
  - γ-quantiles
  - Exponential
  - Joint distributions
  - More distributions

Continuous random variables

Definition 29

A continuous random variable X is a function

$$X : S \to \mathbb{R}$$

where $X(S)$ is uncountable.

The distribution of X is defined by the probability density function $f_X : \mathbb{R} \to \mathbb{R}_0^+$ with the property

$$\int_{-\infty}^{+\infty} f_X(x)\ dx = 1.$$

Measure Theory

Definition 30

- A Borel set of $\mathbb{R}$ is any subset $A \subseteq \mathbb{R}$ which can be represented as countably many unions and intersections of intervals (open, half-open, or closed) on $\mathbb{R}$.
- A function $f : \mathbb{R} \to \mathbb{R}$ is (Borel-)measurable if the preimage of any Borel set also is a Borel set.
- For a measurable function f we denote the Lebesgue integral by $\int f \, d\lambda$.

Example 31 (Examples of measurable functions)

- the characteristic function 1_A of the set A,
- continuous functions, and
- sums and products of measurable functions.

Probability spaces over Borel sets

The set of Borel sets $\mathcal{A}$ is a σ-algebra over $\mathbb{R}$.

A Borel-measurable function f with the properties of a PDF defines the probability space $(\mathbb{R}, \mathcal{A}, P)$ with

$$P : A \mapsto \int f \cdot 1_A \, d\lambda.$$

Especially, P satisfies the Kolmogorov axioms.

Continuous probability spaces

Definition 32

An event is a set $A = \bigcup_k I_k \subseteq \mathbb{R}$ that can be resembled as the union of countably many pairwise disjunct intervals. The probability of A is given as

$$P(A) = \int_A f_X(x)\ dx = \sum_k \int_{I_k} f_X(x)\ dx.$$

The probability of the event $A = \{x\}, x \in \mathbb{R}$ is always 0.

Cumulative distribution functions

The cumulative distribution function of a continuous random variable X is given as

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(X < x) \\ &= \int_{-\infty}^{x} f_X(t)\, dt. \end{aligned}$$

The PDF of X can be obtained from the CDF of X by finding its derivative with respect to x:

$$f_X(x) = \frac{dF_X}{dx}.$$

Intervals

By the fundamental theorem of calculus, the probability of X being in the interval between a and b is given as

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\ dx.$$

Expected values

The expected value of a continuous random variable X is given as

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f_X(x)\ dx.$$

The law of the unconscious statistician still holds in the continuous case:

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x) \cdot f_X(x)\ dx.$$

Uniform

Definition 33 ($X \sim Unif(a, b)$)

A continuous random variable X is uniformly distributed with parameters a and b when X models the outcome of an experiment where all outcomes that lie in the interval $[a, b]$ are equally likely.

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases} . \qquad F_X(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases} .$$

Overview

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

Universality of the Uniform

Let $X \sim F$. Then $F(X) \sim \mathit{Unif}(0,1)$.

Realizations of a random variable of any distribution F with the inverse CDF F^{-1} can be simulated using realizations of a uniformly distributed random variable Y: $F^{-1}(Y) \sim F$.

Normal (Gaussian)

Definition 34 ($X \sim \mathcal{N}(\mu, \sigma^2)$)

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) =: \varphi(x; \mu, \sigma).$$

$$F_X(x) =: \Phi(x; \mu, \sigma).$$

Overview

- $E(X) = \mu$
- $Var(X) = \sigma^2$
- $M_Z(s) = exp(\mu s + \frac{(\sigma s)^2}{2})$

$\mathcal{N}(0,1)$ is the standard normal distribution.

Linear transformation

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then for any $a \in \mathbb{R} \setminus \{0\}$ and $b \in \mathbb{R}$ the random variable

$$Y = aX + b$$

is normally distributed with mean $a\mu + b$ and variance $a^2\sigma^2$.

Standardization

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = \frac{X-\mu}{\sigma}$. Then $Y \sim \mathcal{N}(0,1)$.
The random variable Y is called standardized.

Additivity

Let $X_1, \ldots, X_n$ independent and normally distributed with parameters μ_i, σ_i^2. Then the random variable

$$Z = a_1 X_1 + \cdots + a_n X_n$$

is normally distributed with mean $a_1\mu_1 + \cdots a_n\mu_n$ and variance $a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2$.

Normal approximation to the Binomial

Let $X \sim Bin(n, p)$ with CDF $F_n(t)$. Then

$$F_n(t) \approx \Phi\left(\frac{t - np}{\sqrt{p(1-p)n}}\right)$$

can be used as an approximation if $np \geq 5$ and $n(1-p) \geq 5$.

γ-quantiles

Definition 35

Let X be a continuous random variable with distribution F_x. A number x_γ with

$$F_X(x_\gamma) = \gamma$$

is called γ-quantile of X or the distribution F_X.

Definition 36

For the standard normal z_γ denotes the γ-quantile.

Exponential

Definition 37 ($X \sim Exp(\lambda)$)

A continuous random variable X is exponentially distributed with parameter λ when X models the time between events in a Poisson process.

$$f_X(x) = \lambda e^{-\lambda x}. \qquad\qquad F_X(x) = 1 - e^{-\lambda x}.$$

Overview

- $E(X) = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$
- $M_X(s) = \frac{\lambda}{\lambda - s}, s < \lambda$

Scaling

Let $X \sim Exp(\lambda)$. If $a > 0$, then $Y = aX$ is exponentially distributed with the parameter λ/a.

Memorylessness

The exponential distribution is the only memoryless continuous distribution. Therefore, any continuous random variable X where

$$P(X > y + x | X > x) = P(X > y)$$

holds for all $x, y > 0$ is exponentially distributed.

Waiting for multiple events

Let $X_1, \ldots, X_n$ be independent, exponentially distributed random variables with parameters $\lambda_1, \ldots, \lambda_n$. Then $X = \min\{X_1, \ldots, X_n\}$ is exponentially distributed with parameter $\lambda_1 + \cdots + \lambda_n$.

Exponential approximation to the Geometric

Let $X_n \sim Geom(\lambda/n)$. The distribution of scaled geometrically distributed random variables $Y_n = \frac{1}{n} X_n$ converges to an exponential distribution with parameter λ as $n \to \infty$.

Poisson process

Let $T_1, T_2, \ldots \sim Exp(\lambda)$ i.i.d. that model the time between the $(i-1)$-st and i-th event.
For $t > 0$ we define

$$X(t) = \max\{n \in \mathbb{N} \mid T_1 + \cdots + T_n \leq t\}$$

resembling the number of events that occurred up until time t.
Then $X(t)$ is Poisson-distributed with parameter $t\lambda$.

Joint distributions

Getting marginals

Given a joint distribution $f_{X,Y}$ the marginal distribution f_X can be obtained as follows:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\ dy.$$

Calculating probabilities

Given an event $A \in \mathbb{R}^2$ the probability of A is the area under the probability density function of X:

$$P(A) = \iint\limits_{A} f_{X,Y}(x,y)\ dx\ dy.$$

Finding PDFs

Given a joint CDF $F_{X,Y}$ the joint PDF $f_{X,Y}$ can be obtained as follows:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y).$$

Finding CDFs

Given a joint PDF $f_{X,Y}$ the joint CDF $F_{X,Y}$ can be obtained as follows:

$$F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u,v) \; du \; dv.$$

More distributions

Definition 38 ($X \sim \mathit{Lognormal}(\mu, \sigma^2)$)

A continuous random variable X is logarithmically normal distributed with parameters μ and σ^2 when $Y = \mathit{ln}(X) \sim \mathcal{N}(\mu, \sigma^2)$.

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot \mathit{exp}\left(-\frac{(\mathit{ln}(x) - \mu)^2}{2\sigma^2}\right)$$

for $x > 0$.

Plan I

Inductive Statistics

- Estimators
- Maximum likelihood estimators
- Law of Large Numbers
- Central Limit Theorem
- Confidence intervals
- Hypothesis tests
- Statistical tests

Inductive Statistics

Inductive statistics aims to use measured quantities to draw conclusions about underlying laws.
To generate data n independent copies of an identical experiment modeled by the random variable X are conducted. A measurement resulting from one of these experiments is called a sample.
Each sample is represented by a separate random variable X_i called sample variable.

Estimators

Definition 39

An estimator for parameter θ is a random variable composed of multiple sample variables used to estimate θ.

The bias of an estimator U is given as $E(U - \theta)$.

An estimator U is unbiased for the parameter θ if $E(U) = \theta$ (i.e. its bias is zero).

Definition 40

The sample mean $\bar{X}$ is an unbiased estimator for $E(X)$.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Definition 41

The sample variance S^2 is an unbiased estimator for $Var(X)$.

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.}$$

Definition 42

The mean squared error is a qualitative measure of an estimator U.

$$MSE(U) = E((U-\theta)^2).$$

If U is unbiased, then $MSE(U) = Var(U)$.

An estimator A is more efficient than another estimator B if $MSE(A) < MSE(B)$.

An estimator U is consistent in mean square if $MSE(U) \xrightarrow{n\to\infty} 0$.

Maximum likelihood estimators

Maximum Likelihood Construction is a procedure to construct estimators for parameters of a given distribution. We find the parameter under which the given samples are most likely. In other words, we find the most likely function to explain the given samples.

Given sample variables $\overrightarrow{X} = (X_1, \ldots, X_n)$ and samples $\overrightarrow{x} = (x_1, \ldots, x_n)$, find Maximum-Likelihood estimator for X with parameter θ.

1. construct $L(\overrightarrow{x}; \theta) = f_{\overrightarrow{X}}(\overrightarrow{x}; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta)$, modeling the likelihood that the samples $\overrightarrow{x}$ are described by θ
2. find θ maximizing L, or equivalently $\ln L(\overrightarrow{x}; \theta) = \sum_{i=1}^{n} \ln f_{X_i}(x_i; \theta)$
3. the value for θ maximizing L is a Maximum-Likelihood estimator for θ

Law of Large Numbers

The law of large numbers says that the sample mean of i.i.d. sample variables $\bar{X}$ converges to the actual mean $E(X)$ with probability 1 as the sample size n approaches infinity.

$$P(|\bar{X} - E(X)| \geq \delta) \leq \epsilon$$

for $\delta, \epsilon > 0$ and $n \geq \frac{Var(X)}{\epsilon\delta^2}$.

Central Limit Theorem

The central limit theorem says that the normalized sum of sample values tends towards a standard normal distribution as the sample size approaches infinity even if the original data is not normally distributed.

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{n\to\infty} \mathcal{N}(0,1) \text{ in distribution}$$

for X_i i.i.d..

Equivalently:

$$\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \xrightarrow{n\to\infty} \mathcal{N}(0,1) \text{ in distribution.}$$

De Moivre-Laplace theorem

The De Moivre-Laplace theorem is a special case of the central limit theorem and states that the Normal distribution can be used as an approximation for the Binomial distribution.

Let $X_1, \dots, X_n \sim Bern(p)$ i.i.d. and $H_n = X_1 + \cdots + X_n$. Then

$$H_n^* = \frac{H_n - np}{\sqrt{np(1-p)}} \xrightarrow{n \to \infty} \mathcal{N}(0,1) \text{ in distribution.}$$

Confidence intervals

Often two estimators are used to approach the estimated quantity from both directions.
The two estimators U_1 and U_2 are chosen such that

$$P(U_1 \leq \theta \leq U_2) \geq 1 - \alpha.$$

The probability $1 - \alpha$ is called confidence level.

If for a concrete sample we calculate the estimators U_1 and U_2 and expect $\theta \in [U_1, U_2]$, then we are only wrong with probability α.
$[U_1, U_2]$ is a confidence interval.

Often a single estimator U is used to define the symmetrical confidence interval $[U - \delta, U + \delta]$.

Hypothesis tests

Given sample variables $\overrightarrow{X} = (X_1, \ldots, X_n)$ and sample values $\overrightarrow{x} = (x_1, \ldots, x_n)$ decide whether to accept or reject a hypothesis.

$K = \{\overrightarrow{x} \in \mathbb{R}^n \mid \overrightarrow{x}$ results in rejecting the hypothesis$\}$ is the critical region (or rejection region) of a test.

K is constructed based on the concrete values of the test variable T that is composed of the sample variables.

A test is called one-sided if K is a half-open interval in $T(S)$ and two-sided if K is a closed interval in $T(S)$.

H_0 is the hypothesis to be tested, also called null-hypothesis.
H_1 is the alternative. H_1 is trivial if it is just the negation of H_0.

Errors

- type 1 error or α-error or significance level
  H_0 holds, but $\overrightarrow{x} \in K$

$$\alpha = \sup_{p \in H_0} P_p(T \in K).$$

- type 2 error or β-error
  H_1 holds, but $\overrightarrow{x} \notin K$

$$\beta = \sup_{p \in H_1} P_p(T \notin K).$$

The quality function g describes the probability that a test rejects the null-hypothesis.

$$g(p) = P_p(T \in K).$$

Statistical tests

Characteristics

Statistical tests can be distinguished by the following characteristics:

- **Number of involved random variables**
  Comparison of two random variables with potentially different distributions (two-sample test), or examination of a single random variable (one-sample test)?
  In case of a two sample test:
  - Independence of involved random variables
    Are independent measurements (independence) or related measurements (dependence) taken?
  - Relationships between several random variables
    Regression analysis describes the examination of functional dependencies between random variables, whereas dependency analysis describes the examination of random variables regarding on independence.

- **Formulation of the null hypothesis**
  Which parameters are examined by the test (e.g. expected value or variance), or is tested for a given distribution?
- **Assumptions**
  Which assumptions does the test make regarding independence, distribution, expected value or variance?

Important statistical tests

- Binomial test
- Z-test
- t-test
- two-sample t-test
- χ^2-test

Plan I

Markov chains

Stochastic processes

Markov property

Representations

Probabilities

Hitting times

Stationary distribution

Interlude: Diagonalization

Convergence

Properties

Stochastic processes

Definition 43

A stochastic process is a sequence of random variables $(X_t)_{t \in T}$ that describe the behavior of a system at time t.

If $T = \mathbb{N}_0$, the stochastic process has discrete time. If $T = \mathbb{R}_0^+$, the stochastic process has continuous time.
If X_t is discrete (i.e. its range is countable), the system is said to have a distinct state at time t.

Markov property

Definition 44

A stochastic process fulfills the Markov property if the probability distribution of the states at time $t+1$ solely depends on the probability distribution of states at time t, but not on the states at times $< t$.

This property can be formalized as follows:

$$P(X_{t+1} = j | X_t = i_t, \dots, X_0 = i_0) = P(X_{t+1} = j | X_t = i_t) =: p^t_{i_t j}.$$

Definition 45

A (finite) Markov chain (with discrete time) over the state space $S = \{0, \ldots, n-1\}$ consists of an infinite sequence of random variables $(X_t)_{t \in \mathbb{N}_0}$ with codomain S and the initial distribution q_0 with $q_0^T \in \mathbb{R}^n$. q_0 represents a valid probability mass function (as a row vector) of the random variable X_0.
Farther, the Markov property must hold.

Representations

Definition 46

If the transition probabilities $p_{ij} = P(X_{t+1} = j | X_t = i)$ are constant over time t, the Markov chain is called (time-)homogeneous.

In that case the transition matrix is given as $P = (p_{ij})_{0 \leq i,j < n}$.

The transition diagram is a graph consisting of vertices S and weighted edges represented by the adjacency matrix P.

A concrete instance of the system can be interpreted as a random walk on the transition diagram.

Probabilities

The distribution of a Markov chain can be identified iteratively for larger and larger t:

$$\begin{aligned} q_{t+1} &= q_t \cdot P \\ q_t &= q_0 \cdot P^t \\ q_{t+k} &= q_t \cdot P^k. \end{aligned}$$

Definition 47

q_t is the state vector (or distribution) of the Markov chain at time t.

The entries of P^k refer to the probability of transitioning from state i to state j in exactly k steps:

$$p_{ij}^{(k)} = P(X_{t+k} = j | X_t = i) = (P^k)_{ij}.$$

Hitting times

Definition 48

The hitting time of state j from state i is modeled by the following random variable:

$$T_{ij} = \min\{n \geq 1 \mid X_n = j \text{ given } X_0 = i\}.$$

The expected hitting time is given as

$$\begin{aligned} h_{ij} &= E(T_{ij}) \\ &= 1 + \sum_{k \neq j} p_{ik} h_{kj}. \end{aligned}$$

The probability of reaching state j from state i in arbitrarily many steps is called arrival probability f_{ij}:

$$\begin{aligned} f_{ij} &= P(T_{ij} < \infty) \\ &= p_{ij} + \sum_{k \neq j} p_{ik} f_{kj}. \end{aligned}$$

Definition 49

The random variable $T_i = T_{ii}$ refers to the recurrence time of state i to state i.

The expected recurrence time $h_i = h_{ii}$ and the recurrence probability $f_i = f_{ii}$ are defined analogously to the expected hitting time and the arrival probability.

Stationary distribution

Definition 50

A state vector π with $\pi = \pi \cdot P$ is a stationary distribution of a Markov chain.

A Markov chain does not necessarily converge to a stationary distribution. Convergence depends on the properties of the Markov chain itself and its initial distribution.

Interlude: Diagonalization

For eigenvectors x_i and related eigenvalues λ_i of a matrix A, $A \cdot x_i = \lambda_i \cdot x_i$ holds.

Then for a square matrix A with eigenvectors $x_1, \ldots, x_n$ and related eigenvalues $\lambda_1, \ldots, \lambda_n$, it holds that

$$\begin{aligned} A \cdot \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} &= \begin{bmatrix} \lambda_1 x_1 & \cdots & \lambda_n x_n \end{bmatrix} \\ &= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n \end{bmatrix} . \end{aligned}$$

Let V be the matrix consisting of the eigenvectors of A as column vectors and let Λ be the diagonal matrix consisting of the eigenvalues of A.

Then $V^{-1} \cdot A \cdot V = \Lambda$ is called diagonalization of A.
Conversely, $A = V \cdot \Lambda \cdot V^{-1}$ holds.

Convergence

From the diagonalization of the transition matrix it follows that

$$P^t = V \cdot \Lambda^t \cdot V^{-1}.$$

This can be used to describe the behavior of a Markov chain for $t \to \infty$:

$$\begin{aligned}
&\lim_{t\to\infty} q_t = \lim_{t\to\infty} q_0 \cdot P^t. \\
&\lim_{t\to\infty} P(X_t = j \mid X_0 = i) = \lim_{t\to\infty} P^t(i,j).
\end{aligned}$$

Properties

Certain properties of Markov chains allow us to draw conclusions about its stationary distributions.

Definition 51

A state i is absorbing if $p_{ii} = 1$, that is its vertex in the transition diagram has no outgoing edges.

A state i is recurrent if $f_i = 1$, that is with probability 1 the Markov chain returns to state i.
if conversely $f_i < 1$, the state i is transient.

Definition 52

A Markov chain is irreducible if every state is reachable from every other state with a positive probability if the Markov chain is run for enough steps. Formally:

$$\forall i, j \in S.\ \exists n \in \mathbb{N}.\ p_{ij}^{(n)} > 0.$$

A finite Markov chain is irreducible if and only if its transition diagram is strongly connected.

If a finite Markov chain is irreducible

- $f_{ij} = 1, \forall i, j \in S$;
- h_{ij} exists, $\forall i, j \in S$; and
- there exists a unique stationary distribution π with $\pi(j) = \frac{1}{h_j}, \forall j \in S$.

The Markov chain does not necessarily converge to the stationary distribution (periodicity!).

We now want to examine the periodicity of states.

Definition 53

For a state i define

$$T(i) = \{n \geq 1 \mid P^n(i,i) > 0\}.$$

Then the period of state i is defined as $d_i = gcd(T(i))$.

If a Markov chain is irreducible, all of its states share the same period. This period is then referred to as the period of the Markov chain.

Definition 54

A state i is aperiodic if $d_i = 1$, or equivalently, if $\exists n_0 \in \mathbb{N}.\ \forall n \geq n_0.\ p_{ii}^{(n)} > 0$.

Therefore a state i is aperiodic if and only if the transition diagram has a closed path from i to i with length n for all $n \in \mathbb{N}$ greater some $n_0 \in \mathbb{N}$.

That is state i is surely aperiodic if in the transition diagram

- it has a loop ($p_{ii} > 0$) or
- it is on at least two closed paths P_1 and P_2 whose lengths are coprime.

A Markov chain is aperiodic if all its states are aperiodic.

Definition 55

An ireducible and aperiodic Markov chain is called ergodic.

For every finite ergodic Markov chain it holds independently of its initial distribution q_0 that

$$\lim_{t \to \infty} q_t = \pi$$

where π refers to its unique stationary distribution.

Definition 56

A square matrix A is called stochastic if all its rows sum to one. Every transition matrix P is stochastic.

Additionally, A is called doubly stochastic if also all its columns sum to one.

For every finite ergodic Markov chain whose transition matrix is doubly stochastic its unique stationary distribution assigns the same probability to each state:

$$\pi \equiv \frac{1}{|S|}.$$