Intelligent Data Analysis and Machine Learning

**Probability theory: Cheat sheet**

University of Potsdam, Department of Computer Science, 2019

Lena Jäger

Let X,Y,Z be random variables with range $R_X$, $R_Y$, $R_Z$, respectively.

# 1 Distributions

## 1.1 Discrete random variables

A discrete random variable is a random variable whose range is either finite or countably infinite (e.g., $\mathbb{N}$, $\mathbb{Q}$, $\{1, 0\}$ or $\{1, 2, 3, 4, 5, 6\}$).

1. The probability distribution is defined by a probability mass function $P$ that assigns a probability $P(X = x) \in [0, 1]$ to each $x \in R_X$. Instead of $P(X = x)$, one can also write $P_X(x)$.
   i) The sum of all probabilities must be 1: $\sum_{x \in R_X} P(X = x) = 1$
   ii) The probability that one out of a union of n elementary events $x_1, x_2, \dots x_n$ occurs equals the sum of their individual probabilities: $P(\bigcup_{i=1}^{n} x_i) = \sum_{i=1}^{n} P(x_i)$.
   iii) The probability that one out of a union of disjoints events $A_1, A_2, \dots A_n$ occurs equals the sum of the individual probabilities of each event: $P(\bigcupdot_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$.
   iv) It is not possible that nothing happens: $P(X = \emptyset) = 0$
2. The cumulative probability mass function gives the probability of $X$ being less or equal than a certain number: $cmf(x) = P(X \leq x) = \sum_{y \leq x} P(X = y)$.[1]

## 1.2 Continuous random variables

A continuous random variable is a random variable whose range is uncountably infinite (e.g., the real numbers or an interval on the real numbers).

1. The probability distribution is defined by a densitiy function $p : R_X \to \mathbb{R}_0^+$ where $R_X$ is $\mathbb{R}$ or a closed interval $[a, b]$ on $\mathbb{R}$.
   i) The probability of X taking a value within an interval $[a, b]$ is the definite integral of the density function over this interval (the area under the graph of the density function): $P(X \in [a, b]) = \int_a^b p(x)dx$
   ii) The probability of X taking a specific value $a$ is zero: $P(X = a) = \int_a^a p(x)dx = 0$
   iii) The integral over all possible values of $X$ is 1: $P(X \in \mathbb{R}) = \int_{-\infty}^{\infty} p(x)dx = 1$ if $R_X = \mathbb{R}$ and $P(X \in [a, b]) = \int_a^b p(x)dx = 1$ if $R_X = [a, b]$.

[1] Usually, statistics textbooks use $F(x)$ rather thatn $cmf(x)$. In this summary, I am using $cmf$ to refer to the cumulative probability mass function in order to clearly distinguish it from the cumulative *density* function used for continuous random variables which is usually also referred to by $F(x)$ (see below).

2. The cumulative density function $cdf(x)$ gives the probability of $X$ being less or equal than a certain number $x$:[2]

   i) $cdf(x) = P(X \leq x) = \int_{-\infty}^{x} p(x)dx$ if $R_X = \mathbb{R}$

   ii) $cdf(x) = P(X \leq x) = \int_{a}^{x} p(x)dx$ if $R_X = [a, b]$

# 2 Conditional probabilities and joint distributions

1. Conditional probability: $p(X|Y) = \frac{p(X,Y)}{p(Y)}$
2. Product rule: $p(X, Y) = p(X|Y)p(Y)$
3. Sum rule:
   a) Discrete random variables: $p(x) = \sum_{y} p(x, y)$
   b) Continuous random variables: $p(x) = \int_{-\infty}^{\infty} p(x, y)dy$
4. Bayes' Theorem: $P(X|Y) = \frac{P(X)\,P(Y|X)}{P(Y)}$

# 3 Independence of random variables

1. X,Y are independent $\Leftrightarrow p(X, Y) = p(X)p(Y)$
2. X,Y are independent $\Leftrightarrow p(X|Y) = p(X)$
3. X,Y are independent $\Leftrightarrow p(Y|X) = p(Y)$
4. X,Y are independent $\Leftrightarrow Cov[X, Y] = 0$ (assuming that the covariance exists)

# 4 Conditional independence of random variables

1. X,Y are independent given Z $\Leftrightarrow p(X, Y|Z) = p(X|Z)p(Y|Z)$
2. X,Y are independent given Z $\Leftrightarrow p(Y|X, Z) = p(Y|Z)$
3. X,Y are independent given Z $\Leftrightarrow p(X|Y, Z) = p(X|Z)$

# 5 Expectation

1. Discrete random variables: $\mathbb{E}[X] = \sum_{x \in R_X} x \cdot P(X = x)$
2. Continuous random variables: $\mathbb{E}[X] = \int_{R_X} x \cdot p(x)dx$

*Note that $\mathbb{E}[X]$ may not exist!*

[2] Usually, statistics textbooks use $F(x)$ rather that $cdf(x)$. In this summary, I am using *cdf* to refer to the cumulative density function in order to clearly distinguish it from the cumulative *probability mass* function used with discrete random variables which is usually also referred to by $F(x)$ (see above).

# 6 Variance

1. The variance of a random variable is its mean squared distance from the mean: $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$

   Plugging in the definition of the expectation for continuous and discrete random variables yields:

   a) Discrete random variables: $Var[X] = \sum_{x \in R_X} (x - \mathbb{E}[X])^2 P(X = x)$

   b) Continuous random variables: $Var[X] = \int_{R_X} (x - \mathbb{E}[X])^2 p(x) dx$

2. Verschiebungssatz: $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

3. Linear transformation: $Var[aX + b] = a^2 Var[X] \qquad (a, b \in \mathbb{R})$

4. The variance of the sum of two random variables equals the sum of their variances plus 2 times their covariance: $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$

5. The variance of the sum of $n$ random variables $X_1, X_2, ..., X_n$ is the sum of their variances plus 2 times the covariance of all possible pairs:
   $Var[\sum_{i=1}^{n} X_i] = \sum_{i,j=1}^{n} Cov[X_i, X_j] = \sum_{i=1}^{n} Var[X_i] + \sum_{i \neq j} Cov[X_i, X_j] = \sum_{i=1}^{n} Var[X_i] + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Cov[X_i, X_j]$

# 7 Covariance

1. If $\mathbb{E}[Y], \mathbb{E}[Y]$ and $\mathbb{E}[XY]$ exist, the covariance of X and Y is defined as $Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])]$

2. The covariance of a random variable with itself equals its variance: $Cov[X, X] = Var[X]$

3. Verschiebungssatz: $Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

4. Biliniarity of the covariance:

   i) $Cov[aX + b, cY + d] = ac \cdot Cov[X, Y] \qquad (a, b, c, d \in \mathbb{R})$

   ii) $Cov[X, (eY + f) + (gZ + h)] = eCov[X, Y] + gCov[X, Z] \qquad (e, f, g, h \in \mathbb{R})$

5. Symmetry of the covariance: $Cov[X, Y] = Cov[Y, X]$

6. Positive semi-definiteness of the covariance: $Cov[X, Y] \geq 0$

7. Cauchy-Schwarz inequality: $|Cov[X, Y]| \leq \sqrt{Var[X]} \cdot \sqrt{Var[Y]}$

# 8 Correlation

1. The Pearson correlation coefficient $\rho$ normalizes the covariance to the interval $[-1, 1]$:

   $\rho(X, Y) = \frac{Cov[X,Y]}{\sqrt{Var[X]} \cdot \sqrt{Var[Y]}}$

# 9 Specific univariate distributions

## 9.1 Discrete distribution

1. Discrete uniform distribution

   *Notation:* $X \sim Unif(a, b)$ or $U(a, b)$

   *Example:* Fair dice: Finite set of outcomes which are all equally likely.

*Parameters:* $a \in \mathbb{Z}$, $b \in \mathbb{Z}$ with $b \geq a$
Interpretation of the parameter in the example: $a$ is the smallest and $b$ the largest possible value that X can take.

*Support:* {a,a+1, a+2, ... b-2, b-1, b}

*Probability mass function:* $P(X = x) = \frac{1}{b-a+1}$

*Cumulative probability mass function:* $P(X \leq x) = \frac{\lfloor x \rfloor - a + 1}{b-a+1}$

*Expectation:* $\mathbb{E}[X] = \frac{a+b}{2}$

*Variance:* $Var[X] = \frac{(b-a+1)^2-1}{12}$

2. Bernoulli distribution

*Notation:* $X \sim Ber(p)$

*Example:* Whether or not a head was achieved in a single coin toss

*Parameters:* $p \in (0, 1)$
Interpretation of the parameter in the example: $p$ is the probability of a head

*Support:* {0,1}

*Probability mass function:* $P(X = x) = \begin{cases} p & \text{if x=1} \\ 1-p & \text{if x=0} \end{cases}$

*Cumulative probability mass function:* $P(X \leq x) = \begin{cases} 0 \text{ if } x < 0 \\ p-1 \text{ if } 0 \leq x < 1 \\ 1 \text{ if } x \geq 1 \end{cases}$

*Expectation:* $\mathbb{E}[X] = p$

*Variance:* $Var[X] = p(1-p)$

3. Binomial distribution

*Notation:* $X \sim Binom(n, p)$

*Example:* Number of heads in an $n$-times repeated coin toss

*Parameters:* $n \in \mathbb{N}$, $p \in (0, 1)$
Interpretation of the parameters in the example: $n$ is the number of coin tosses; $p$ is the probability of a head in a single coin toss

*Support:* $\{0, 1, 2, ...n\}$

*Probability mass function:* $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

*Cumulative probability mass function:* $P(X \leq x) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$

*Expectation:* $\mathbb{E}[X] = np$

*Variance:* $Var[X] = np(1-p)$

4. Geometric distribution

*Notation:* $X \sim Geom(p)$

*Example:* The number of Bernoulli trials needed to get one success.[3]

*Support:* $\mathbb{N}^+$

*Parameters:* $p \in (0, 1]$
Interpretation of the parameter in the example: $p$ is the probability of a success in one Bernoulli trial.

*Probability mass function:* $P(X = x) = (1-p)^{x-1} p$

[3]There are two version of the geometric distribution. In the version not presented here, the number $Y = X - 1$ of *failures* before the first success is modeled. In that version, the support of Y is $\mathbb{N}_0$.

*Cumulative probability mass function:* $P(X \leq x) = 1 - (1-p)^x$

*Expectation:* $\mathbb{E}[X] = \frac{1}{p}$

*Variance:* $Var[X] = \frac{1-p}{p^2}$

5. Poisson distribution

   *Notation:* $X \sim Pois(\lambda)$

   *Example:* Number of events occurring in a fixed time interval or a given area. The average rate in which the events occur is assumed to be known and each event occurs independently of the time since the last event.

   *Support:* $\mathbb{N}_0$

   *Parameters:* $\lambda \in \mathbb{R}^+$
   Interpretation of the parameter in the example: $\lambda$ is the rate with which the event occurs (i.e., the average number of times the event occurs within one time interval).

   *Probability mass function:* $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$

   *Cumulative probability mass function:* $P(X \leq x) = e^{-\lambda} \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!}$

   *Expectation:* $\mathbb{E}[X] = \lambda$

   *Variance:* $Var[X] = \lambda$

## 9.2 Continuous distribution

1. Continuous uniform distribution

   *Notation:* $X \sim Unif(a, b)$

   Explanation: The uniform distribution is the distribution with maximal entropy (i.e., uncertainty about the outcome).

   *Support:* $x \in [a, b]$

   *Parameters:* $a, b \in (-\infty, +\infty)$
   Interpretation of the parameters:$a$ ist the minimum and $b$ is the maximum.

   *Probability density function:* $p(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

   *Cumulative probability density function:* $P(X \leq x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b) \\ 1 & \text{for } x \geq b \end{cases}$

   *Expectation:* $\mathbb{E}[X] = \frac{1}{2}(a + b)$

   *Variance:* $Var[X] = \frac{1}{12}(b - a)^2$

2. Normal distribution (also called Gaussian distribution)

   *Notation:* $X \sim N(\mu, \sigma^2)$

   *Usage:* Central limit theorem

   *Special case:* Standard normal distribution: $N(0, 1)$

   *Support:* $x \in \mathbb{R}$

   *Parameters:* $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_0^+$
   Interpretation of the parameters: $\mu$ is the mean (i.e., the expectation), the mode and the median of the distribution; $\sigma^2$ is the variance. The larger $\sigma$ is, the wider the distribution.

   *Probability density function:* $P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

*Cumululative density function:* $P(X \leq x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$

*Expectation:* $\mathbb{E}[X] = \mu$

*Variance:* $Var[X] = \sigma^2$

3. Student's t-distribution

   *Notation:* $X \sim t_n$

   *Usage:* t-test;

   *Relation to the Normal distribution:* for $n \to \infty$, the t-distribution and the standard normal distribution are identical.

   *Support:* $x \in \mathbb{R}$

   *Parameters:* $n \in \mathbb{N}^+$
   Interpretation of the parameter: $n$ are the degrees of freedom.

   *Probability density function:* $p(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$ Note that $\Gamma$ refers to the $\Gamma$-function.

   *Cumululative density function:* ugly

   *Expectation:* $\mathbb{E}[X] = 0$ if $n > 1$ and undefined if $n = 1$.

   *Variance:* $Var[X] = \sigma^2$

4. Beta distribution

   *Notation:* $X \sim Beta(\alpha, \beta)$

   *Usage in Bayesian statistics:* Conjugate prior to several likelihood distributions: Bernoulli, Binomial, Geometric etc. distributions.

   *Support:* $x \in (0, 1)$

   *Parameters:* $\alpha, \beta \in \mathbb{R}^+$
   Interpretation of the parameters: If $\alpha = \beta$, the distribution is symmetric. Depending on the values of *alpha* and *beta*, the distribution can be either convex or concave.

   *Probability density function:* $p(x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}$ where $B(p,q) := \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 u^{p-1}(1-u)^{q-1}\, du$

   *Cumululative density function:* $P(X \leq x) = \begin{cases} 0 & \text{if } x \leq 0, \\ \frac{1}{B(p,q)} \int_0^x u^{p-1}(1-u)^{q-1} du & \text{if } 0 < x \leq 1, \\ 1 & \text{if } x > 1 \end{cases}$

   *Expectation:* $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$

   *Variance:* $Var[X] = Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

5. Gamma distribution[4]

   *Notation:* $X \sim Gamma(\alpha, \beta)$

   *Usage in Bayesian statistics:* Conjugate prior to several likelihood distributions: Poisson, Exponential, Normal etc. distributions.

   *Support:* $x \in \mathbb{R}_0^+$

   *Parameters:* $\alpha \in \mathbb{R}^+$, $\beta \in \mathbb{R}^+$
   Interpretation of the parameters: $\alpha$ is the shape parameter and $\beta$ the rate parameter

   *Probability density function:* $p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

   *Cumululative density function:* $p(x) = \frac{1}{\Gamma(\alpha)} \gamma(\alpha, \beta x)$

   *Expectation:* $\mathbb{E}[X] = \frac{\alpha}{\beta}$

   *Variance:* $Var[X] = \frac{\alpha}{\beta^2}$

[4] There is also another parametrization of the Gamma distribution. The one introduced here is the one usually used in Bayesian statistics.

# 10 Specific multivariate distributions

## 10.1 Continuos distributions

1. Multivariate normal distribution (i.e., joint distribution of n normally distributed variables)

   *Notation:* $N(\mu, \Sigma)$ *$\mu$ is a vector of length $n$ and $\Sigma$ is a non-negative symmetric $n \times n$ matrix*

   *Usage:* Central limit theorem

   *Support:* $x \in \mathbb{R}^n$

   *Parameters:* $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$. $\Sigma$ is symmetric and $\sigma_{ij} \geq 0 \forall i, j \in \{1, ...n\}$
   Interpretation of the parameters: $\mu$ is the vector of means of the $n$ random variables $X_i$; $\Sigma$ is the covariance matrix of the random variables where $\sigma_{ij} = Cov[X_i, X_j]$.

   *Probability density function:* $p(x) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$

   *Expectation:* $\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ ... \\ \mathbb{E}[X_n] \end{pmatrix} = \mu$

   *Covariance:* $Cov[X] = \Sigma$

   *Variance:* $Var[X_i] = \sigma_{ii}$
   The variances are on the diagonal of the covariance matrix because Var[$X_i$]=Cov[$X_i$,$X_i$]