Analysis of System Performance

IN2072

Chapter 1 – Statistics

Dr. Alexander Klein
Prof. Dr.-Ing. Georg Carle

Chair for Network Architectures and Services
Department of Computer Science
Technische Universität München
http://www.net.in.tum.de

Topics

- Random Variable
- Probability Space
- Discrete and Continuous RV
- Frequency Probability(Relative Häufigkeit)
- Distribution(discrete)
- Distribution Function(discrete)
- PDF,CDF
- Expectation/Mean, Mode,
- Standard Deviation, Variance, Coefficient of Variation
- p-percentile(quantile), Skewness, Scalability Issues(Addition)
- Covariance, Correlation, Autocorrelation
- Visualization of Correlation
- Examples

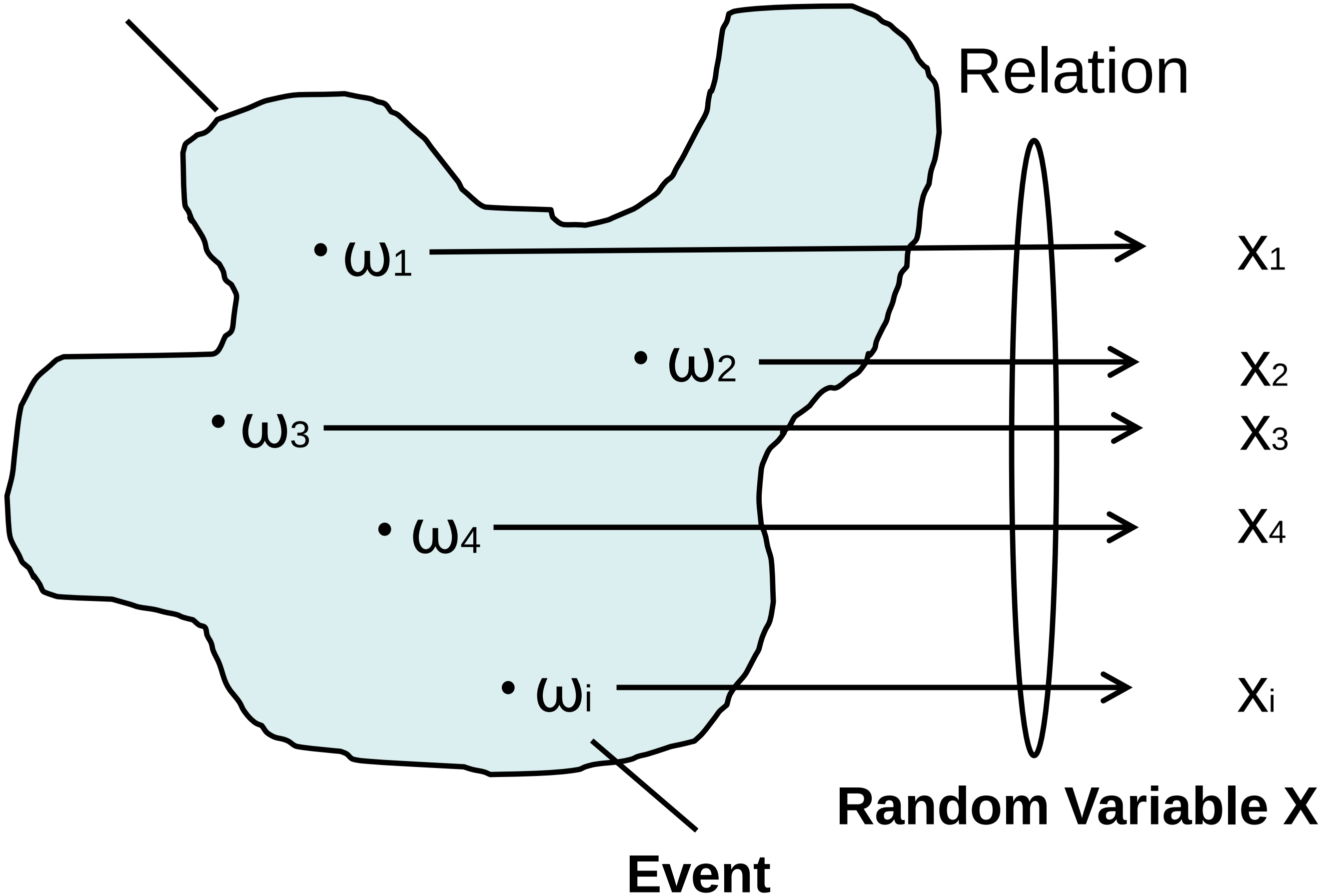
Classic definition of probability

The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible.

\- Pierre-Simon Laplace, A Philosophical Essay on Probabilities

Random Variable

Probability Space (Ereignisraum) $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \ldots, \omega_i\}$

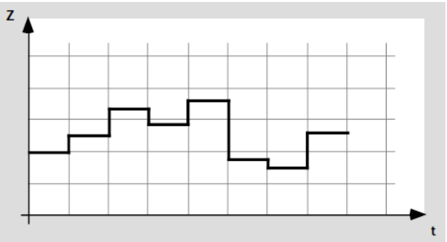


Statistics Fundamentals

- **Discrete Random Variable:**

Countable

  - Example: Flipping of a coin
    - ω_1={head-0}, ω_2={tail-1} => $\Omega = \{\omega_1, \omega_2\}$
    - X e {0, 1}



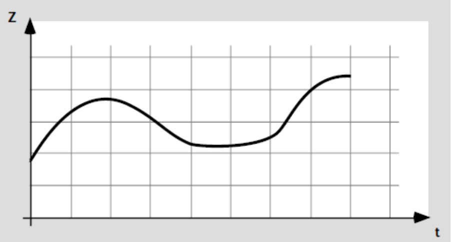
  - Example: Rolling two dice

    - ω_1={2}, ω_2={3}, … , ω_{11}={12} => $\Omega = \{\omega_1, \omega_2, \dots , \omega_{11}\}$
    - X e {2, 3, 4, …, 12}

- **Continuous Random Variable:**

Uncountable

  - Example: Round Trip Time
    - T e {5ms, 200ms}
    - ω_1={t<10ms}, ω_2={10ms≤t<20ms}, ω_3={t≥20ms} => $\Omega = \{\omega_1, \omega_2\}$
  - Example: Sensed Interference Level



Discrete or not discrete

Statistics Fundamentals

- **Frequency Probability** / Law of large numbers (Relative Häufigkeit)

  - Number of random experiments
    - n total number of trials
    - X_i event or characteristic of the outcome
    - n_i number of trials where the event X_i occurred

$$h(X_i) = \frac{n_i}{n} \qquad 0 \le h(X_i) \le 1 \qquad \sum_i h(X_i) = 1$$

Vollständigkeits-relation

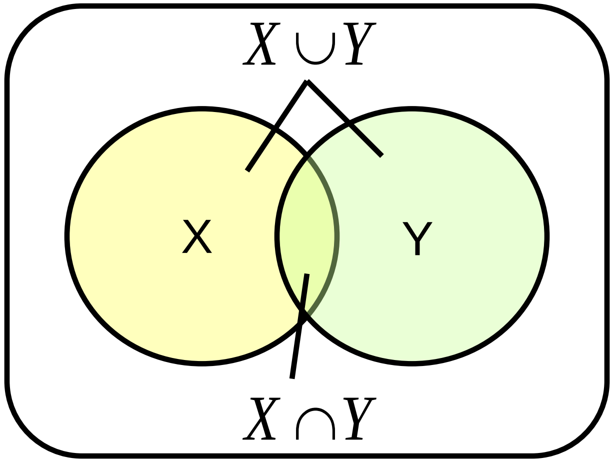
$$P(X_i) = \lim_{n \to \infty} \frac{n_i}{n} \qquad 0 \le P(X_i) \le 1 \qquad \sum_i P(X_i) = 1$$

X_i disjoint

Statistics Fundamentals

- **Vollständiges Ereignissystem**

$$P(Y) = \sum_{i=1}^{N} P(X_i)$$



- **Verbundereignis**

$$P(X \cap Y) = P(X,Y) = P(Y,X)$$

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$

- **Bedingte Wahrscheinlichkeit**

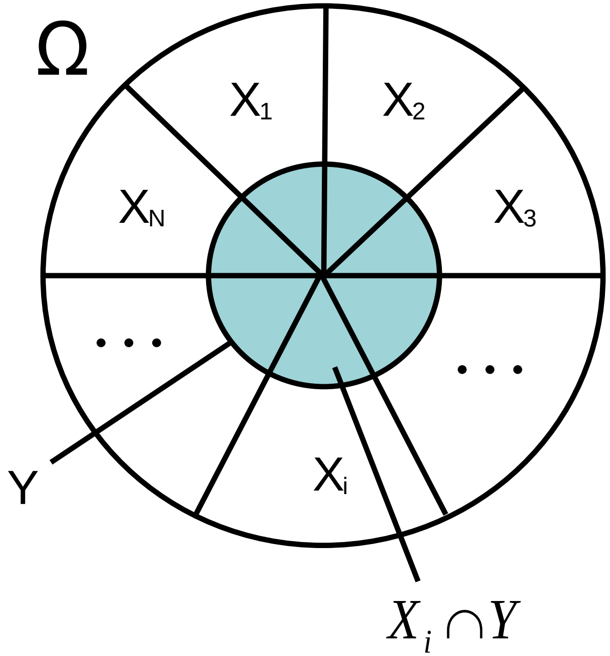
$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)} \Rightarrow P(X \mid Y) \geq P(X,Y)$$

- **Statistische Unabhängigkeit**

$$P(X \mid Y) = P(X) \quad \vee \quad P(X,Y) = P(X)P(Y)$$

Statistics Fundamentals

- **Vollständiges Ereignissystem**
  - $$P(Y) = \sum_{i=1}^{N} P(X_i, Y)$$



- **Bayes Function**
  - $$P(X_i \mid Y) = \frac{P(Y \mid X_i) \cdot P(X_i)}{P(Y)} = \frac{P(Y \mid X_i) \cdot P(X_i)}{\sum_{k=1}^{N} P(Y \mid X_k) \cdot P(X_k)}$$

Statistics Fundamentals

- **Distribution** (Verteilung)

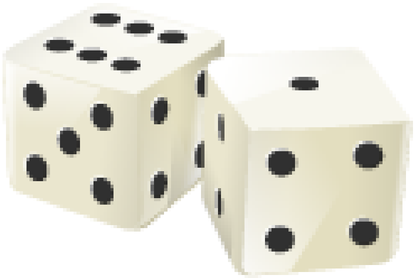
  X – discrete random variable

  - Function $x(i) = P(X = i)$, $i = 0,1,2,\ldots,X_{max}$ (Distribution)
    - – $x(i)$ e $[0,1]$
    - – $\sum_{i=0}^{X_{max}} x(i) = 1$ (Vollständigkeitsrelation)

  - Example:
    Rolling two dice
    - $\omega_1=\{2\}, \omega_2=\{3\}, \ldots, \omega_{11}=\{12\}$ => $\Omega = \{\omega_1, \omega_2, \ldots, \omega_{11}\}$
    - $X \in \{2, 3, 4, \ldots, 12\}$

Statistics Fundamentals

- Example: Throwing two dice



Sample Space (Ereignisraum)

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

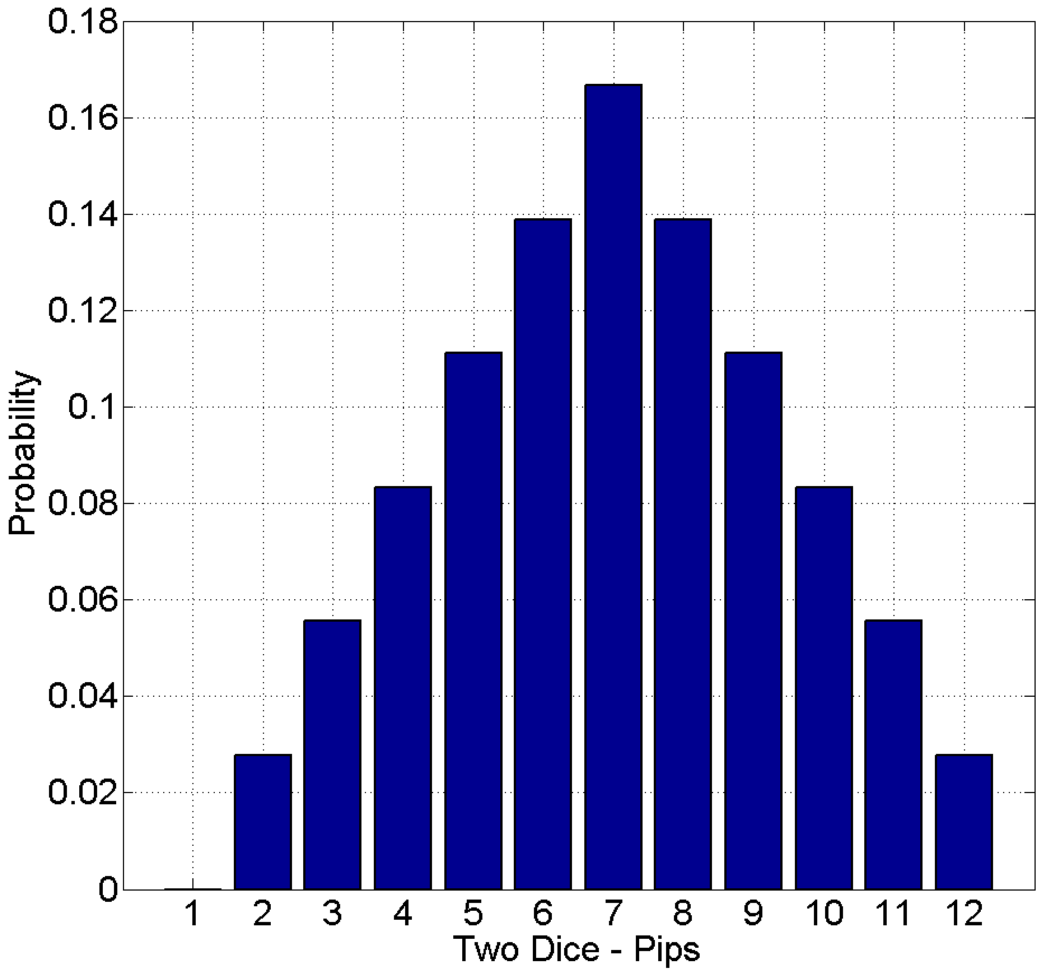
Statistics Fundamentals

- Example: Throwing two dice

Sample Space (Ereignisraum)

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Statistics Fundamentals

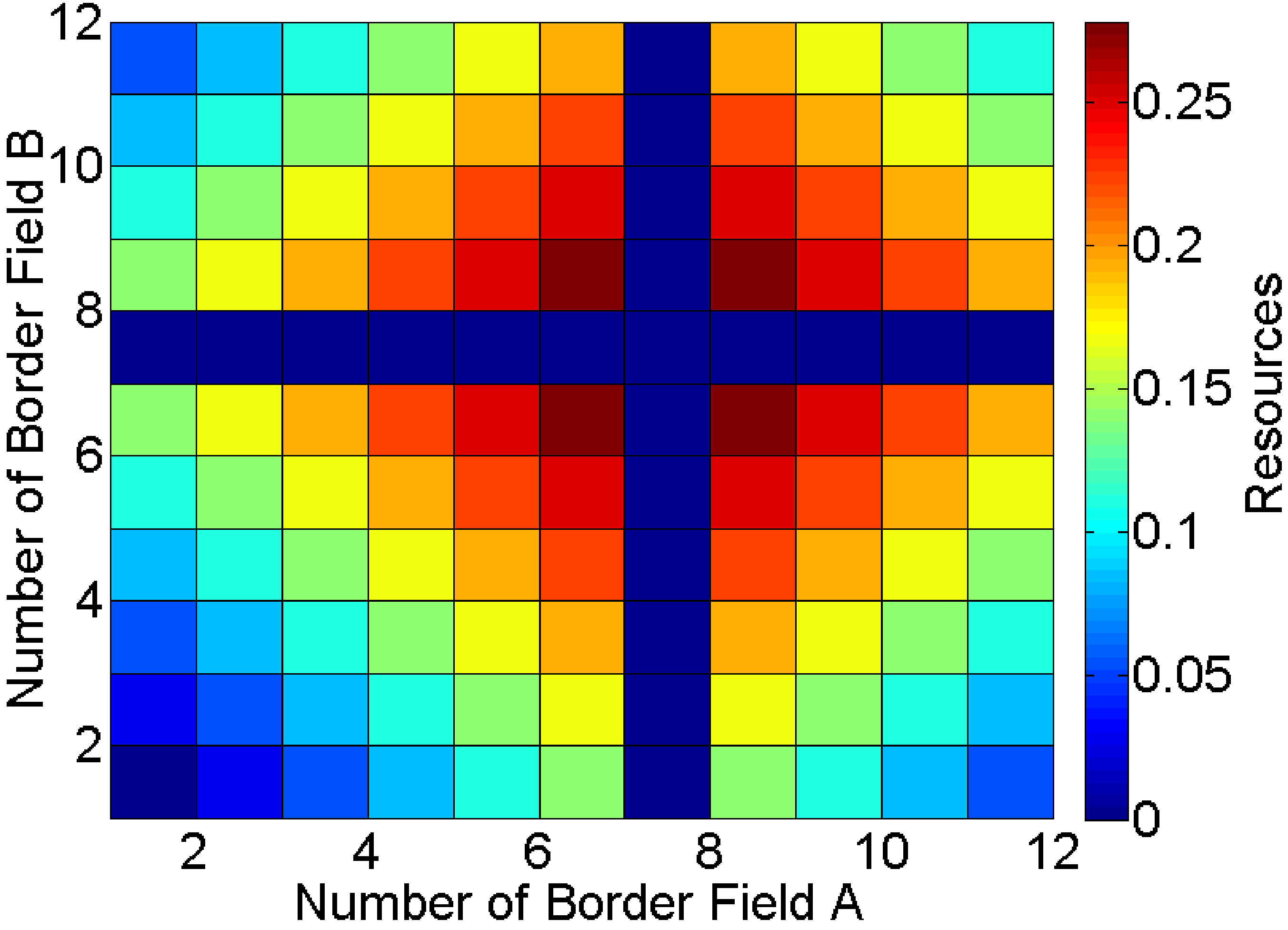


Distribution

Palour Game: Die Siedler von Catan

- Rules:
  - Players are only allowed to build along borders of a field
  - Players roll two dice
  - If the sum of the dice corresponds to the number of the field, the player gets the resources from this field
- Question
  - Where is the best place for a building?

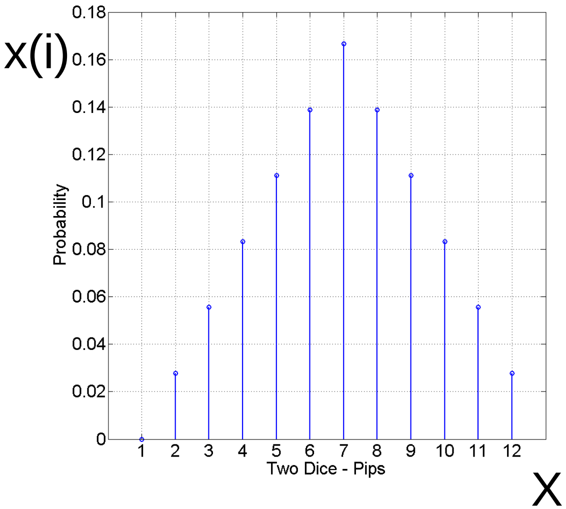
Statistics Fundamentals



9
8
7
6
5
4
3
2
1
Prison

Are all fields reached with the same probability?

- **Probability Mass Function** (Verteilung)
  - Discrete random variable X
  - i value of the random variable X
  - x(i) probability that the outcome of random variable X is i
  - $x(i) = P\{X = i\}, \quad i = 0,1,\ldots, X_{\max}$ (Distribution)
  - $\sum_{i=0}^{X_{\max}} x(i) = 1$ (Vollständigkeitsrelation)

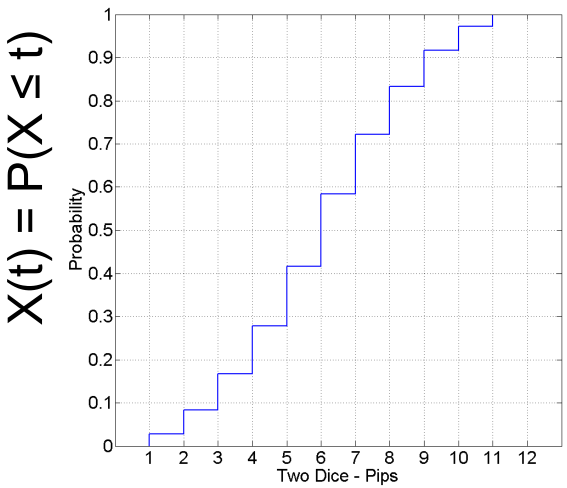


Probability Mass Function

- **Cumulative Distribution Function** (Verteilungsfunktion)

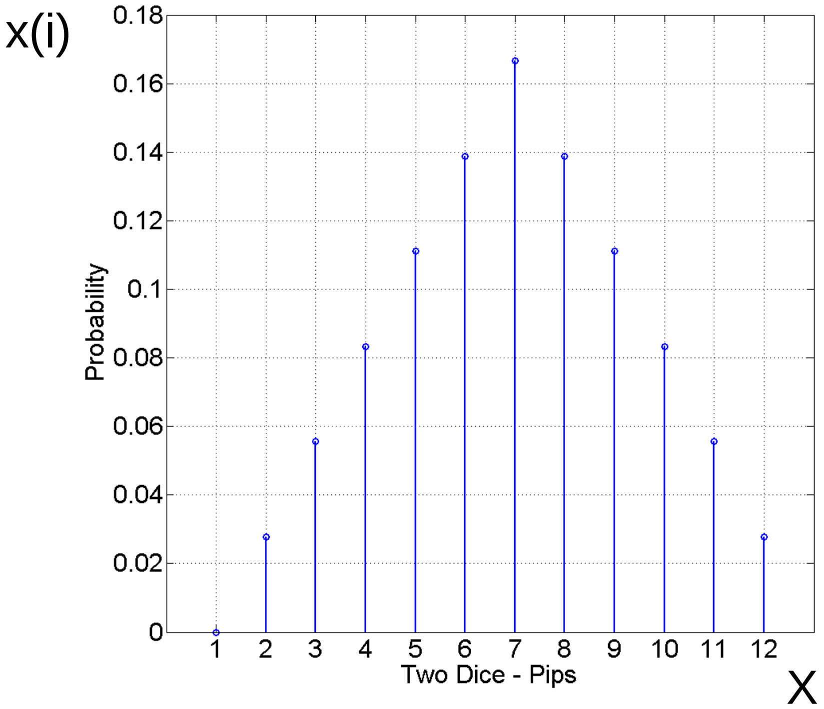
$$X(t) = P\{X \le t\}$$

- $t_1 < t_2 \Longrightarrow X(t_1) \le X(t_2)$ (monotony)
- $t_1 < t_2 \Longrightarrow P\{t_1 < X \le t_2\} = X(t_2) - X(t_1)$
- $X(-\infty) = 0 \quad \wedge \quad X(\infty) = 1$
- $X^c(t) = 1 - X(t) = P\{X > t\}$

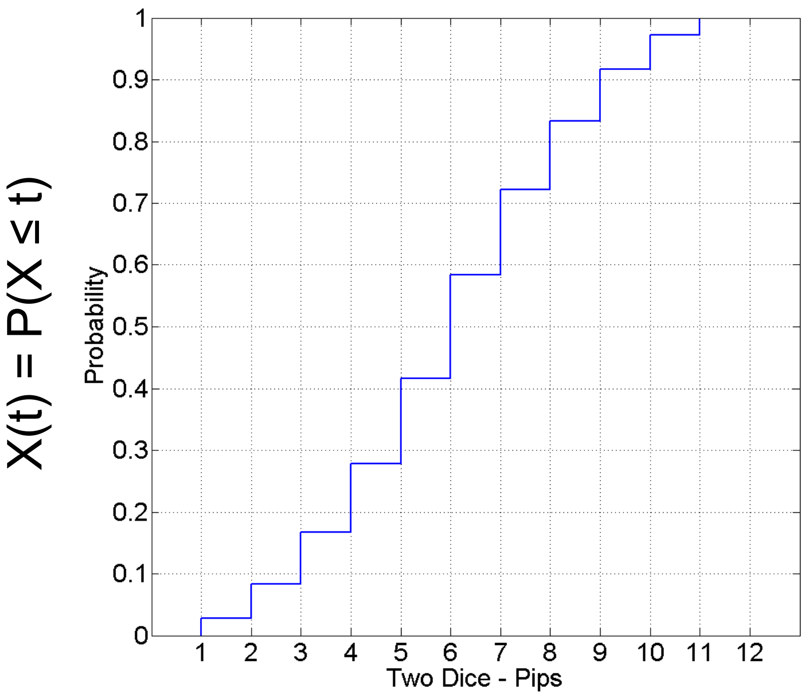


Cumulative Distribution Function

Difference between probability mass function and cumulative distribution function

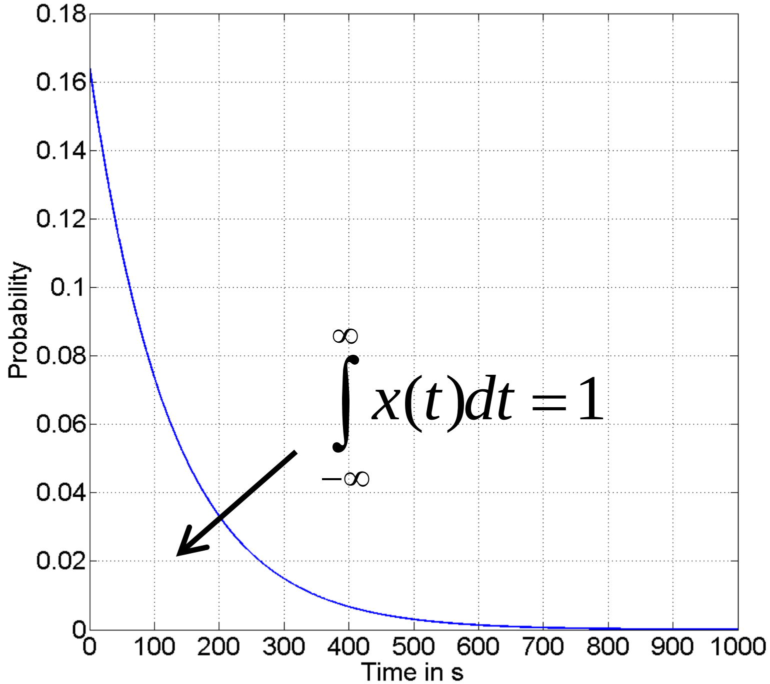


Probability Mass Function (Verteilung)



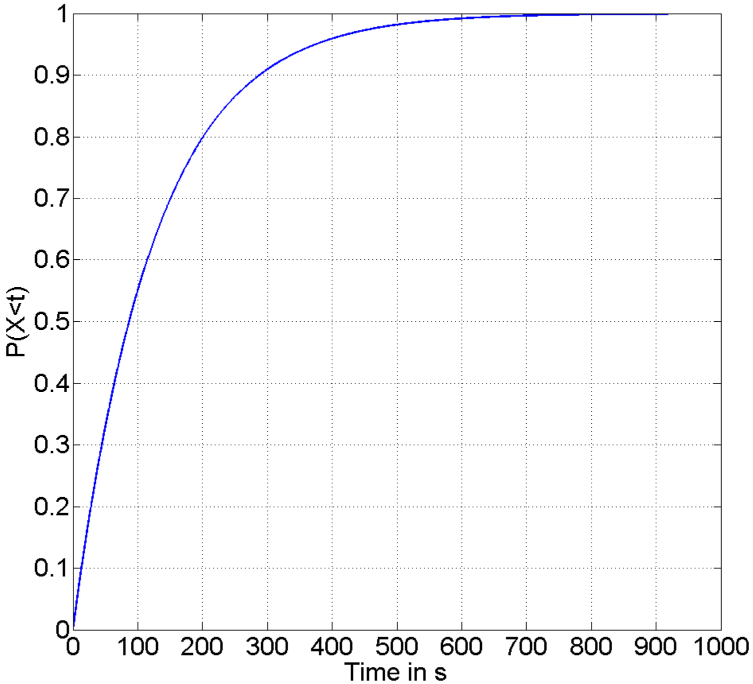
Cumulative Distribution Function (Verteilungsfunktion)

Continuous random variable



Probability Density Function
(Verteilungsdichtefunktion)

$$x(t) = \frac{d}{dt} X(t)$$



Cumulative Density Function

$$X(t) = \int_{-\infty}^{t} x(t)dt$$

Sum of Two Random Variables (Convolution/Faltung)

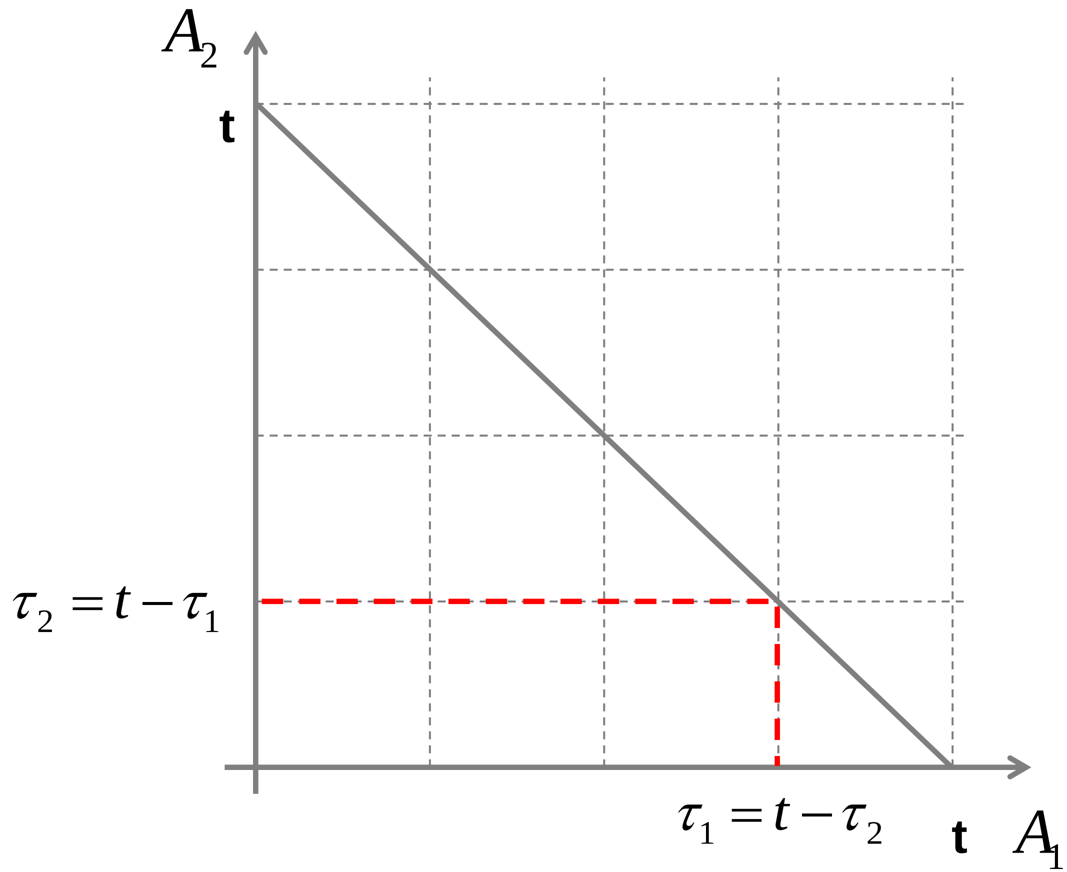
- Discrete
  - Random variables X_1, X_2
  - Distribution $Y = X_1 + X_2$

$$P(Y=i) = y(i) = (x_1 * x_2)(i) = \sum_j x_1(j) \cdot x_2(i-j)$$

- Continuous
  - Random variables A_1, A_2
  - Probability density function $A = A_1 + A_2$

$$a(t) = (a_1 * a_2)(t) = \int_{-\infty}^{t} a_1(\tau) \cdot a_2(t-\tau) d\tau$$

Sum of Two Random Variables (Convolution/Faltung)



$$A(t) = \int_{\tau_1+\tau_2 \le t} a(\tau_1,\tau_2)\,d\tau_1 d\tau_2 = \int_{\tau_1=0}^{t} \left(\int_{\tau_2=0}^{t-\tau_1} a(\tau_1,\tau_2)\,d\tau_2 \right) d\tau_1$$

Difference of Two Random Variables

- Discrete
  - Random variables X_1, X_2
  - Distribution $Y = X_1 - X_2$

$$P(Y = i) = y(i) = (x_1 * x_2)(i) = \sum_j x_1(j) \cdot x_2(i - j)$$

- Continouous
  - Random variables A_1, A_2
  - Probability density function $A = A_1 - A_2$

$$a(t) = (a_1 * a_2)(t) = \int_{-\infty}^{t} a_1(t + \tau) \cdot a_2(\tau) d\tau$$

Maximum of Random Variables

- Random variables A_i
- Distribution $A = \max(A_0, A_1, \ldots, A_{k-1})$
- $k = 2: a(t) = a_1(t) \cdot A_2(t) + a_2(t) \cdot A_1(t), A(t) = A_1(t) \cdot A_2(t)$
- $A(t) = \prod_{0 \le i < k} A_i(t)$

Minimum of Random Variables

- Random variables A_i
- Distribution $A = \max(A_0, A_1, \ldots, A_{k-1})$
- $k = 2: a(t) = a_1(t) \cdot A_2^c(t) + a_2(t) \cdot A_1^c(t), A(t) = A_1^c(t) \cdot A_2^c(t)$
- $A^c(t) = \prod_{0 \le i < k} A_i^c(t)$ ⇔ $A(t) = 1 - \prod_{0 \le i < k} (1 - A_i(t))$

Statistics Fundamentals

- **Expectation** (Erwartungswert)
  - X : Probability density function
  - g(x) : Function of random variable X

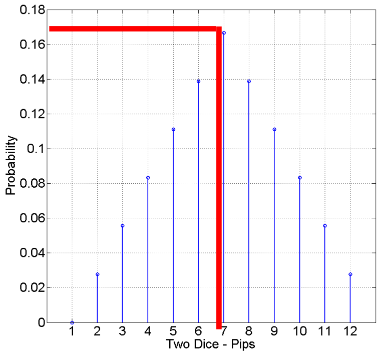
$$E[g(X)] = \int_{-\infty}^{\infty} g(t) \cdot x(t)\,dt$$

- **Mean** (Mittelwert einer Zufallsvariablen)

$$m_1 = E[X] = \int_{-\infty}^{\infty} t \cdot x(t)\,dt$$

- **Mode** (Outcome of the random variable with the highest probability)

$$c = Max(x(t))$$



Statistics Fundamentals

- Gewöhnliche Momente einer Zufallsvariablen
  - $$g(X) = X^k \Longrightarrow m_k = E\left[X^k\right] = \int_{-\infty}^{\infty} t^k \cdot x(t)\,dt, \qquad k = 0,1,2,\ldots$$

- Central moment (Zentrales Moment)
  - Variation of the random variable in respect to its mean

$$g(X) = (X - m_1)^k$$

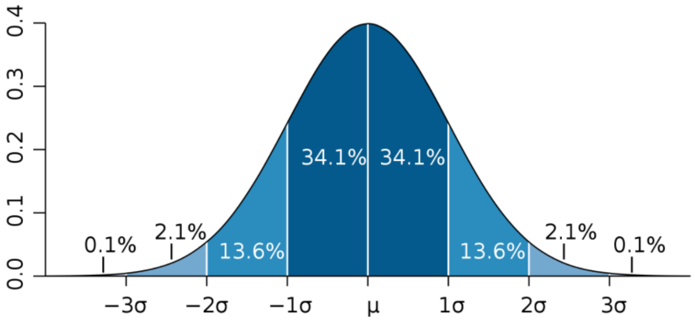
$$\Longrightarrow \mu_k = E\left[(X - m_1)^k\right] = \int_{-\infty}^{\infty} (t - m_1)^k \cdot x(t)\,dt, \qquad k = 0,1,2,\ldots$$

  - Special Case (k=2):

$$\mu_2 = E\left[(X - m_1)^2\right] = VAR[X]$$

Statistics Fundamentals

- **Standard deviation** (Standardabweichung)
  - $$\sigma_X = \sqrt{VAR[X]}$$



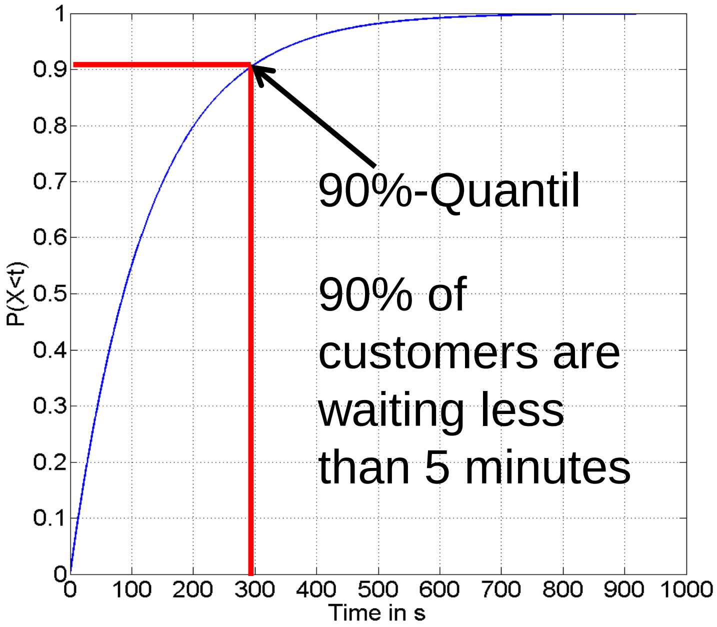
- **Coefficient of variation** (Variationskoeffizient)
  - $$c_X = \frac{\sigma_X}{E[X]}, \qquad E[X] > 0$$
  - The coefficient of variation is a normalized measure of dispersion of a probability distribution
  - It is a dimensionless number which does not require knowledge of the mean of the distribution in order to describe the distribution

Picture taken from Wikipedia

- **p-percentile t_p** (p-Quantil)

  A percentile is the value of a variable below which a certain percent of observations fall

  - VDF $F : R \rightarrow (0,1)$ (bijective)
  - $F(x) = P(X < x) = p$
  - $F^{-1}(x) = \inf\{x \in R : p \leq F(x)\}$
  - Special Case:
    - Median 0.5-percentile
    - Upper percentile 0.75-percentile
    - Lower percentile 0.25-percentile
  - Typical Use Case:
    - QoS in networks
      (e.g. 99.9%-percentile of the delay)



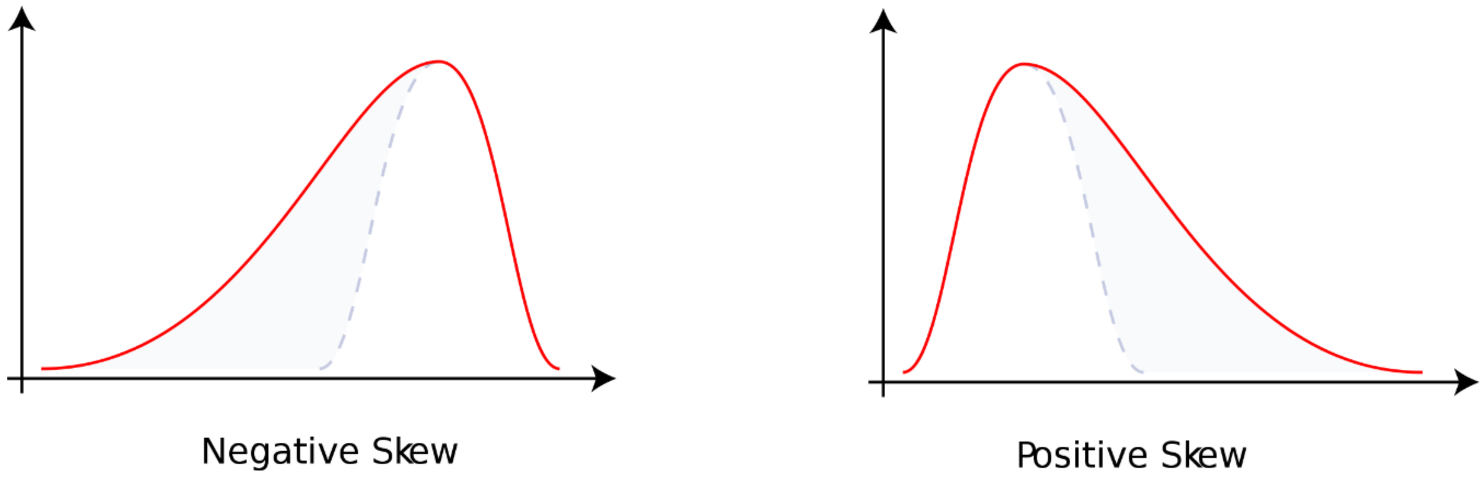
Cumulative Density Function

- **Skewness** (Schiefe)

  Skewness describes the asymmetry of a distribution

  - v < 0 : The left tail of the distribution is longer (linksschief)
    => Mass is concentrated in the right
  - v > 0 : The right tail of the distribution is longer (rechtsschief)
    => Mass is concentrated in the left

$$\upsilon_X = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}$$



Picture taken from Wikipedia

Statistics Fundamentals

Scalability Issues

- Multiplication of a random variable X with a scalar s
  - $Y = s \cdot X$
  - $E[Y] = s \cdot E[X]$
  - $VAR[Y] = s^2 \cdot VAR[X]$
- Addition of two random variables X and Y
  - $Z = X + Y$
  - $E[Z] = E[X] + E[Y]$
  - $VAR[Z] = VAR[X] + VAR[Y]$ (only if A and B independent)

Statistics Fundamentals

Covariance

Covariance is a measure which describes how two variables change together

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

- Special Case: $Cov(X,X) = VAR[X]$
- Other Characteristics:
  - $Cov(X,a) = 0$
  - $Cov(X,Y) = Cov(Y,X)$
  - $Cov(aX,bY) = abCov(X,Y)$
  - $Cov(X+a, Y+b) = Cov(X,Y)$

Correlation function

Correlation function describes how two random variable tend to derivate from their expectation

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{VAR(X) \cdot VAR(Y)}}$$

- Characteristics:
  - $Y = X$ ⇨ $Cor(X,Y) = 1$ (Maximum positive)
  - $Y = -X$ ⇨ $Cor(X,Y) = -1$ (Maximum negative)
  - $Cor(X,Y) > 0$ Both random variable tend to have either high or low values (difference to their expectation)
  - $Cor(X,Y) < 0$ The random variables differ from each other such that one has high values while the other has low values and vice versa (difference to their expectation)

Autocorrelation (LK 4.9)

- Autocorrelation is the cross-correlation of a signal with itself. In the context of statistics it represents a metric for the similarity between observations of a stochastic process. From a mathematical point of view, autocorrelation can be regarded as a tool for finding repeating patterns of a stochastic process.

Definition:

- Correlation of two samples with distance k from a stochastic process X is given by:

$$\Rightarrow \quad Cor(X,Y) \quad \text{with} \quad Y_i = X_{i+j}$$

Use case:

- Test of random number generators
- Evaluation of simulation results (c.f. Batch-Means)

Example:

00101110101001101100010011101010100011

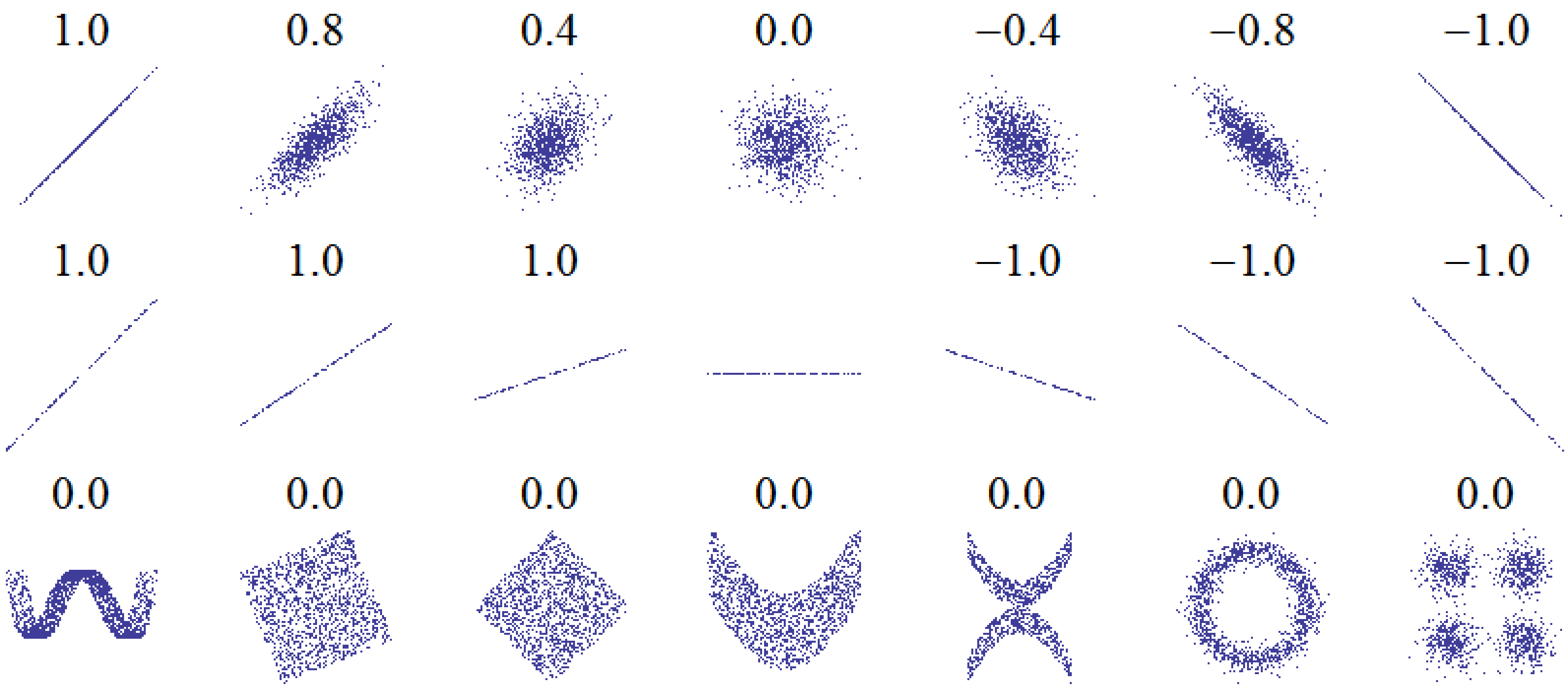
00101110101001101100010011101010100011

Random 

Autocorrelation Lag 4

Visualization of Correlation

Example: Two random variables X and Y are plotted against each other

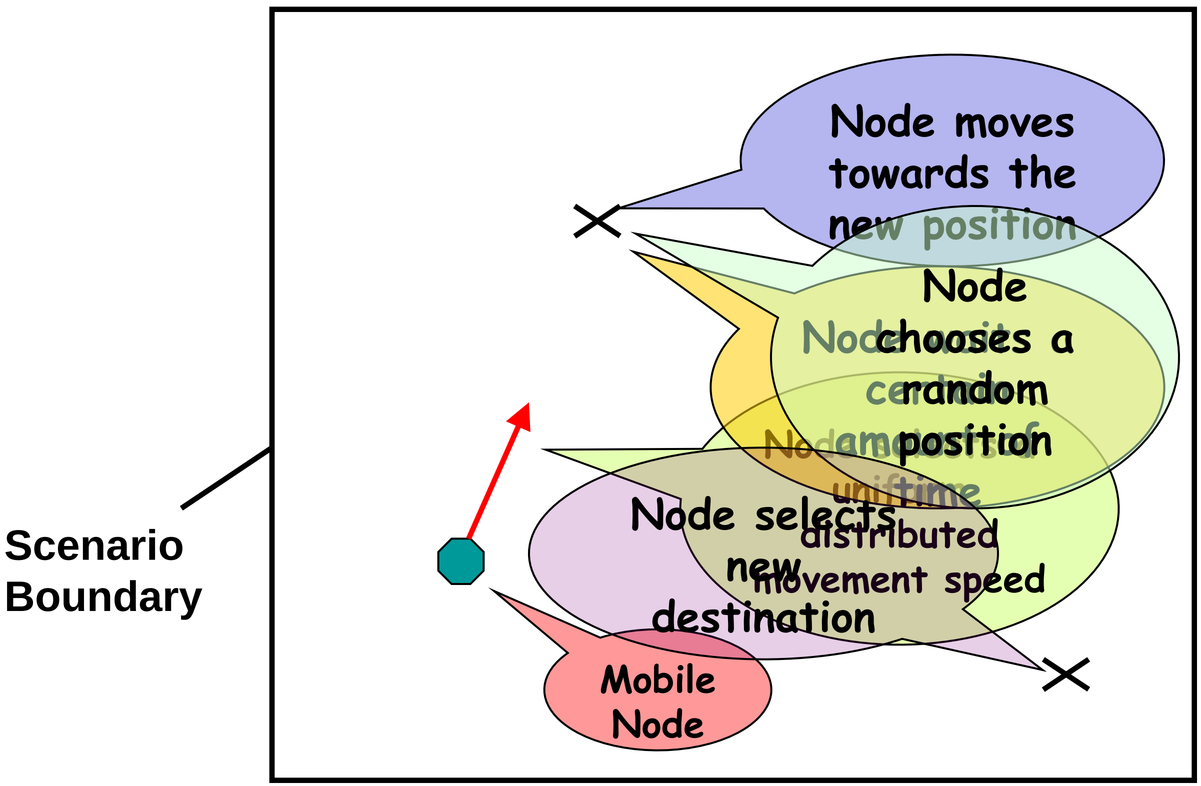


Picture taken from Wikipedia

❑ Impact of correlation (1/2)

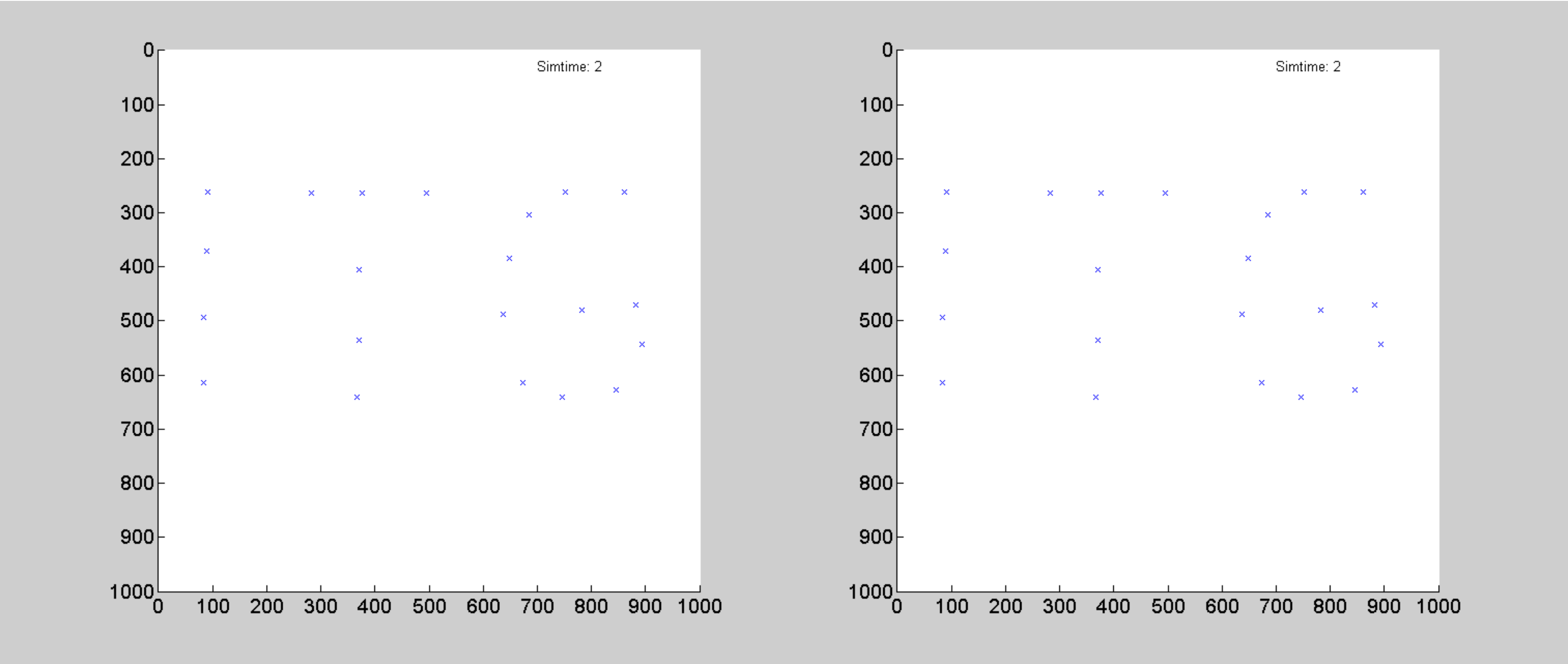
Example: Random Waypoint mobility model

Algorithm



Impact of correlation (2/2)

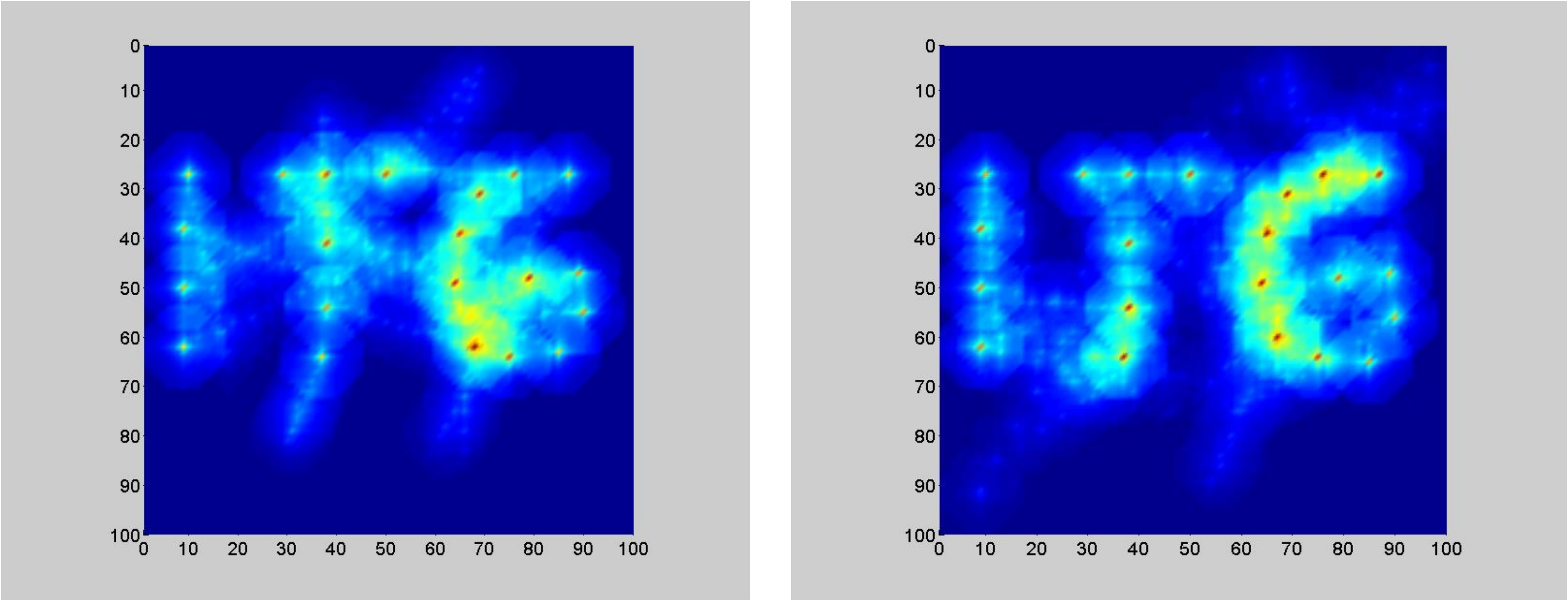
Example: Random Waypoint mobility model



Uncorrelated next position selection

Correlated next position selection

- Mobility Example

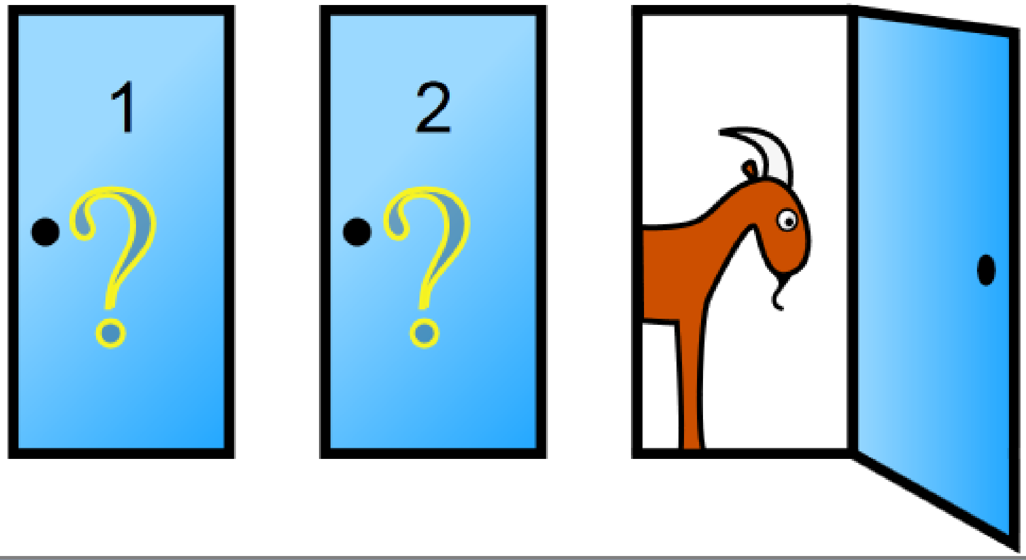


Statistics Fundamentals

- Monty Hall Problem – (also known as the goat problem)
  - American game show „Let‘s make a deal“ adopted in Germany „Geh auf Ganze“

- Game rules:
  - Behind one door is a price
  - Behind the other doors is the goat / Zonk (It is assumed that the candidate is not interested in neither the goat nor the Zonk)
  - Candidate may choose one door
  - Game master will open one door after the decision of the candidate and will offer the candidate the choice to choose a different door.

Should the candidate change his/her decision?



Statistics Fundamentals

- Definition: RV Z=i: „Zonk/Goat is behind door i“
  - P(Z=i)=1/n (Laplace)
- Definition: RV C=i: „Candidate has chosen door i“
  - P(C=i)=1/n (Laplace)
- Z an C are independent
  - $P(Z=i \wedge C=i)=P(Z=i) \cdot P(C=i)$
- Definition: ZV O=i: „Door i was opened“
  - P(O=i|Z=i)=0 „The winning door was not opened“
  - $P(Z=i \wedge O=i) = P(O=i|Z=i) \cdot P(Z=i) = 0$ (Bayes)
  - $P(Z=i \wedge O \neq i) = P(Z=i) - P(Z=i \wedge O=i) = P(Z=i)$ (Totale Wahrscheinlichkeit)
- P(O=i|C=i)=0 „The selected door will NOT be opened“
  - $P(C=i \wedge O=i) = P(O=i|C=i) \cdot P(C=i) = 0$ (Bayes)
  - $P(C=i \wedge O \neq i) = P(C=i) - P(C=i \wedge O=i) = P(C=i)$ (Totale Wahrscheinlichkeit)

Statistics Fundamentals

- Win probability if the player does not change his selection

$$P(Z=i \mid C=i \wedge O \neq i) = \frac{P(Z=i \wedge (C=i \wedge O \neq i))}{P(C=i \wedge O \neq i)}$$

$$= \frac{P(Z=i \wedge C=i)}{P(Z=i)} = \frac{\frac{1}{n} \cdot \frac{1}{n}}{\frac{1}{n}} = \frac{1}{n}$$

- Win probability if player changes his/her selection

$$P(Z=i \mid C \neq i \wedge O \neq i) = \frac{1 - P(Z=i \wedge (C=i \wedge O \neq i))}{n-2}$$

$$= \frac{1-\frac{1}{n}}{n-2} = \frac{n-1}{n \cdot (n-2)}$$

Will Rogers phenomenon (1)

- Revenues per salesman of company HuiSoft for two consecutive years, in k€:

2010		2011	
Bielefeld	München	Bielefeld	München
5000	5000		
6000	10000		
7000	15000		
	20000		
μ=6000	μ=12500		

Will Rogers phenomenon (2)

- Will Rogers (1879–1935),
  American comedian and philosopher
- Named after one of his jokes:

Frage: Wenn die 10% dümmsten Saarländer nach Rheinland-Pfalz ziehen, was passiert dann?

Antwort: In beiden Bundesländern steigt der IQ an.

- (originally with Oklahomans and Californians…)
- Lesson:
  - Will Rogers phenomena are ubiquitous,
  - yet can be difficult to spot
  - …even for the authors themselves!
  - **Warning – it's a sword that cuts both ways:**
    Sometimes looking at the details is better, sometimes looking at the aggregated numbers makes more sense (as in the sales example)

Simpson Paradox (1)

- Universität Eschweilerhof discriminates against female students!
- Let’s see what faculties are the most sexist ones:

	Applications				Acceptance rate	
Faculty	female	acc.	male	acc.	female	male
Engineering	10	8	80	50	80%	63%
CS	5	4	60	40	80%	67%
Philosophy	80	20	40	10	25%	25%
Law	30	15	40	10	50%	25%
Total	125	47	220	110	(←significant numbers)	
Acc. rate		37.6%		50.0%		

- None of them!? How can that be?
  - Women applied at faculties with more competition

Simpson Paradox (2)

- So who is right? Should the university be punished?
  - The women's rights activists? After all, 37.6% vs. 50% is significant – and dividing the total number into faculties simply introduces a bias into the picture.
  - The university? After all, not a single faculty does actually discriminate against women (in fact, most discriminate against men).
- Answer: In *this* case, the university is right
  - A student applies at a specific faculty that he or she chooses herself
  - A student does not apply at university and lets the university choose the faculty
- **Lesson:**
  - Simpson Paradox is more ubiquitous than you would think,
    yet can be difficult to spot …even for the authors themselves!
  - **Warning – it's a sword that cuts both ways:**
    Sometimes looking at the details makes more sense (as in this case), sometimes looking at the aggregated numbers is better.

Simpson Paradox (3)

Ärzte in Hansistan: Jahreseinkommen vor Steuern

	Schnibbler		Tröster		Knochenflicker	
	Flach-land	Berg-land	Flach-land	Berg-land	Flach-land	Berg-land
Anzahl Ärzte	200	40	20	20	30	100
davon über 200 000 Piepen	60	8	16	2	24	70
Anteil Best-verdiener in %	30	20	80	10	80	70

Wo verdienen Ärzte besser – im Flachland oder im Bergland? Im Detail deutet alles aufs Flachland hin.

Ärzte in Hansistan: Jahreseinkommen vor Steuern

	Alle Ärzte	
	Flachland	Bergland
Anzahl	250	160
davon über 200 000 Piepen	100	80
Anteil Best-verdiener in %	40	50

Im Gesamtüberblick scheint das Bergland jedoch vorne zu liegen.

Important terms German/English

Verteilung	distribution
Verteilungsfunktion	distribution function cumulative distribution function (cdf)
Verteilungsdichtefunktion	probability density function (pdf)
p-Quantil	p-percentile, p-quantile
Faltung	convolution
Erwartungswert	mean
Moment	moment
Zentrales Moment	central moment
Varianz	variance
Standardabweichung	standard deviation
Variationskoeffizient	coefficient of variation
Schiefe	skewness
Kovarianz	covariance
Korrelation	correlation
Empirisches Moment Stichprobenmoment	sample moment
Empirischer Mittelwert Stichprobenmittelwert	sample mean
Empirische Varianz Stichprobenvarianz	sample variance
Empirische Schiefe Stichprobenschiefe	sample skewness
Empirische Kovarianz Stichprobenkovarianz	sample covariance
Empirische Autokovarianz Stichprobenautokovarianz	sample autocovariance
Empirische Autokorrelation Stichprobenautokorrelation	sample autocorrelation

Important distributions and their characteristics

Distribution - Continuous

- **Uniform distribution**: $RV \ X \sim U(a,b)$ (LK 8.3.1)
  - Density function: $f(x) = \frac{1}{b-a}, X \in [a;b]$
  - Range: $[a;b]$
  - Distribution function: $F(x) = \frac{x-a}{b-a}$
  - Expectation: $E(X) = \frac{a+b}{2}$
  - Variance: $VAR(X) = \frac{(b-a)^2}{12}$
  - Generation: $U \sim U(0,1), X = a + (b-a)U$

Distribution - Continuous

- **Triangle distribution (1/4)**: $RV \;\; X \sim triang(a,b,c)$ (LK 8.3.15)

  - Density function:

$$f(x) = \begin{cases} \dfrac{2\cdot(x-a)}{(b-a)\cdot(c-a)} & if \;\; a \le x \le c \\ \dfrac{2\cdot(b-x)}{(b-a)\cdot(b-c)} & if \;\; c \le x \le b \\ 0 & otherwise \end{cases}$$

  - Distribution function:

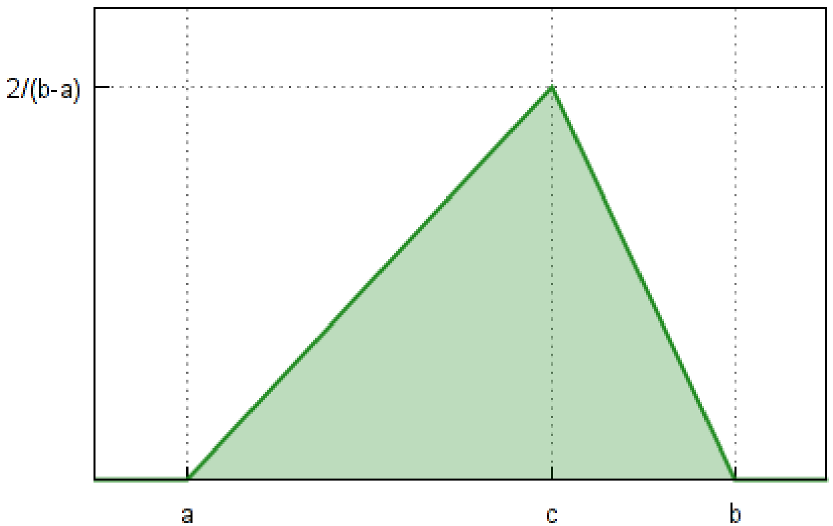
$$f(x) = \begin{cases} 0 & if \;\; x < a \\ \dfrac{(x-a)^2}{(b-a)\cdot(c-a)} & if \;\; a \le x \le c \\ 1-\dfrac{(b-x)^2}{(b-a)\cdot(b-c)} & if \;\; c \le x \le b \\ 1 & if \;\; b < x \end{cases}$$

Distribution - Continuous

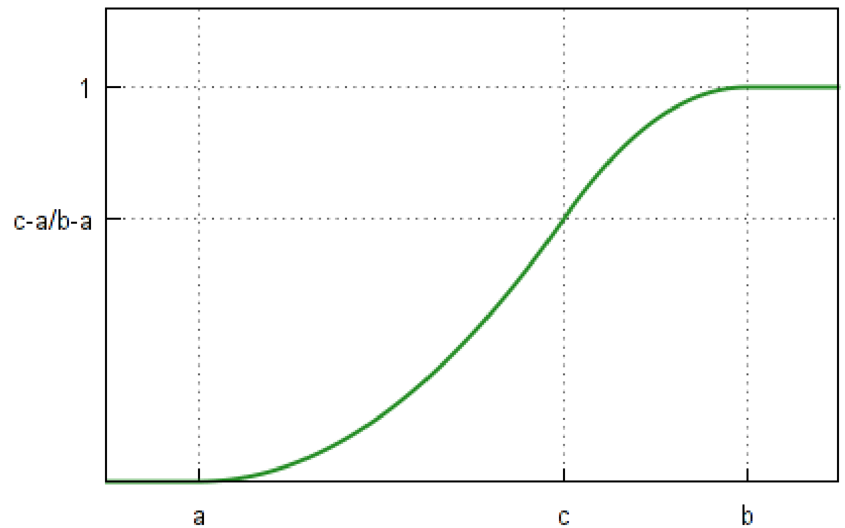
- **Triangle distribution (2/4):** $RV \;\; X \sim triang(a,b,c)$ (LK 8.3.15)
  - Use case: Project management / business simulations where only the minimum, maximum and mode are known
  - Mode c
  - Range $[a;b]$
  - Expectation: $$E(X) = \frac{a+b+c}{3}$$
  - Variance: $$VAR(X) = \frac{(a^2+b^2+c^2-ab-ac-bc)}{18}$$

- **Triangle distribution (3/4):** $RV \;\; X \sim triang(a,b,c)$ (LK 8.3.15)
  - Generation: Inversion

$$U \sim U(0,1),\, X = \begin{cases} a + \sqrt{U(b-a)\cdot(c-a)} & 0 < U < F(c) \\ b - \sqrt{(1-U)\cdot(b-a)\cdot(b-c)} & F(c) < U < 1 \end{cases}$$



Probability Density Function

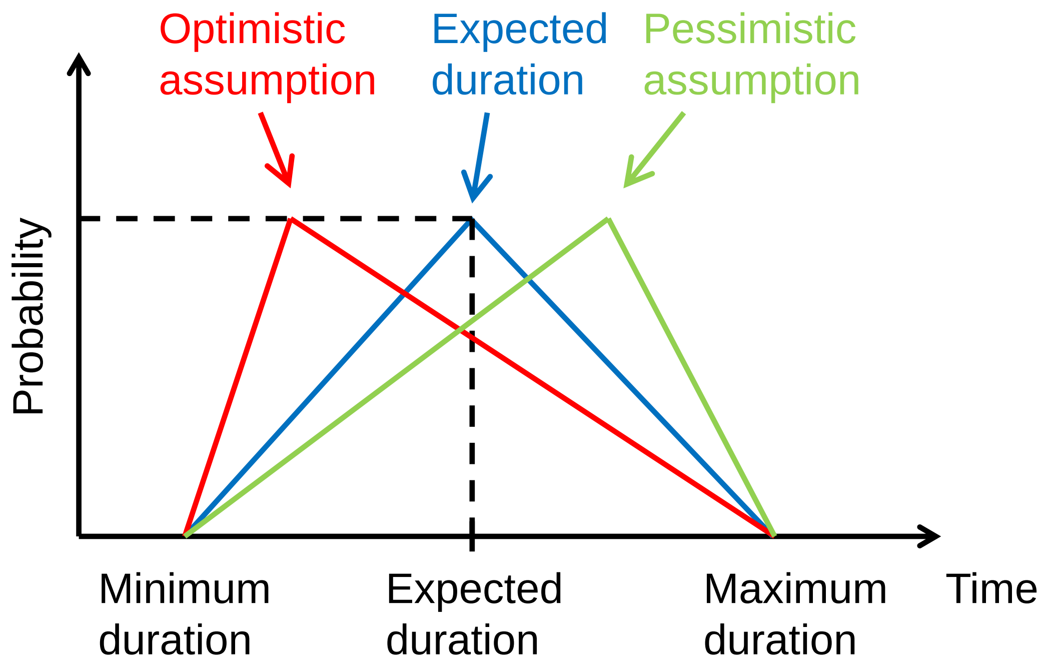


Cumulative Density Function

Distribution - Continuous

- **Triangle distribution (4/4):** $RV\ \ X \sim triang(a,b,c)$ (LK 8.3.15)

  Use case: risk management / project management



Distribution - Continuous

- **Normal distribution(1/4):** $RV \ \ X \sim N(\mu,\sigma^2)$ (LK 8.3.6)
  - Density function: $f(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot e^{-\left(\frac{(x-\mu)^2}{2 \cdot \sigma^2}\right)}$
  - Distribution function: $F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$
  - Range: $\left]-\infty;\infty\right[$
  - Mode: μ
  - Expectation: $E(X) = \mu$
  - Variance: $VAR(X) = \sigma^2$
  - Scalability: $X \sim N(0,1) \Rightarrow (\mu + \sigma X) \sim N(\mu,\sigma^2)$

Distribution - Continuous

- **Normal distribution(2/4):** $RV\ X \sim N(\mu, \sigma^2)$ (LK 8.3.6)
  - Generation Accept-Reject
    - Two independent random variables $U_1, U_2 \sim U(0,1)$
    - $V_i = 2U_i - 1$
    - $W = V_1^2 + V_2^2$
    - Algorithm:

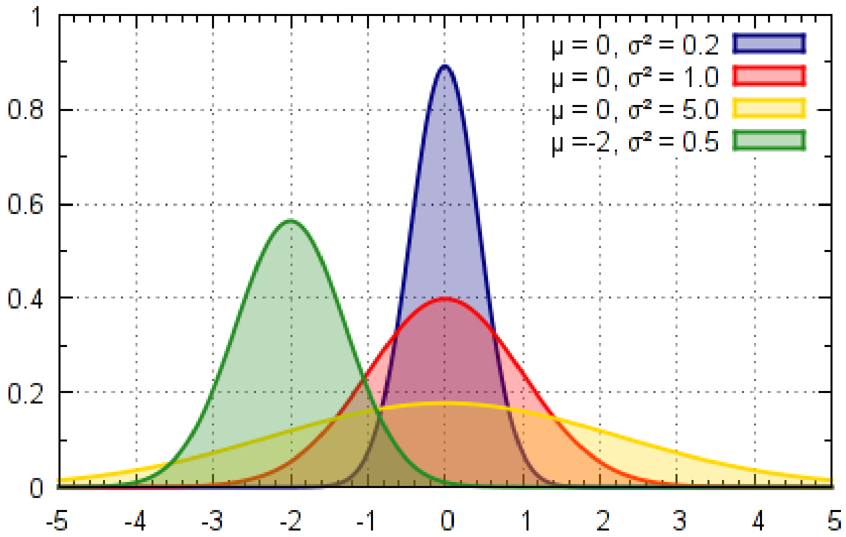
Accept if $W \leq 1$

$$Y = \sqrt{\frac{-2\ln W}{W}}\ ,\quad X_1 = V_1 \cdot Y\ ,\quad X_2 = V_2 \cdot Y$$

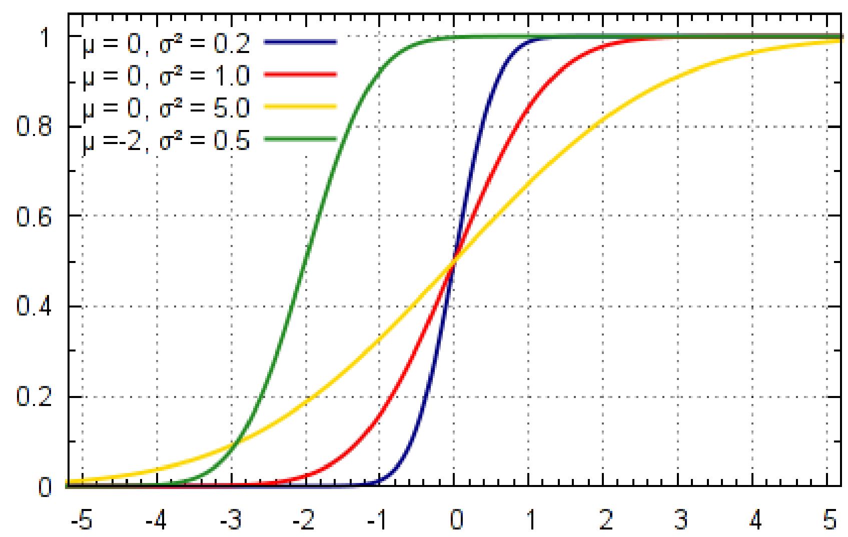
Reject otherwise

Distribution - Continuous

- **Normal distribution(3/4):** $RV\ \ X \sim N(\mu, \sigma^2)$ (LK 8.3.6)



Probability Density Function



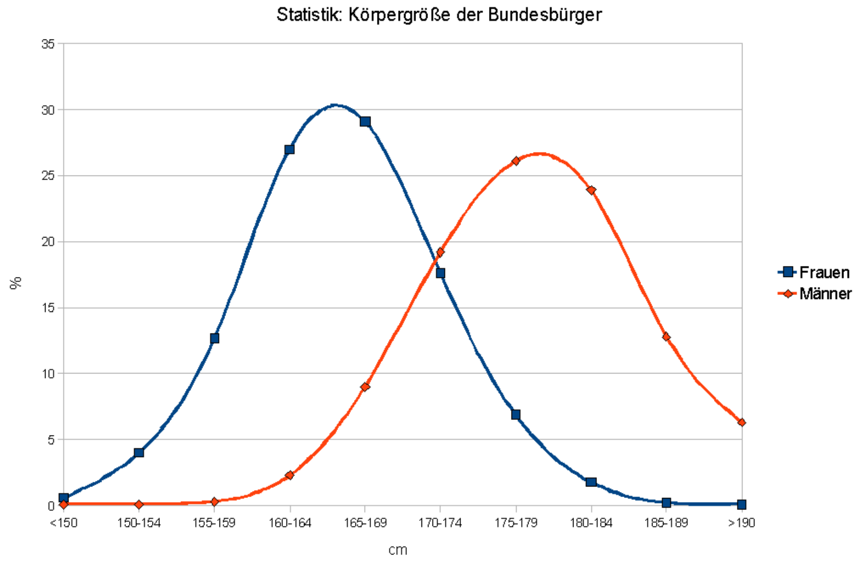
Cumulative Density Function

Distribution - Continuous

- **Normal distribution(4/4):** $RV\ \ X \sim N(\mu, \sigma^2)$ (LK 8.3.6)

  Use case: distribution of errors / sizes (nature)

Körpergröße	Frauen	Männer
<150 cm	0,6 %	0,1 %
150–154 cm	4 %	0,1 %
155–159 cm	12,7 %	0,3 %
160–164 cm	27,0 %	2,3 %
165–169 cm	29,1 %	9,0 %
170–174 cm	17,6 %	19,2 %
175–179 cm	6,9 %	26,1 %
180–184 cm	1,8 %	23,9 %
185–189 cm	0,2 %	12,8 %
≥ 190 cm	<0,1 %	6,3 %



Körpergröße der Deutschen **Statistik des** Sozio-oekonomischen Panels (SOEP)**, aufbereitet durch** statista.org

Distribution - Continuous

- **Lognormal distribution(1/3):** $RV\ \ X \sim LN(\mu,\sigma^2)$ (LK 8.3.7)

  - Special property of the lognormal distribution

    if $Y \sim N(\mu,\sigma^2) \Rightarrow e^Y \sim LN(\mu,\sigma^2)$

  - Range: $[0,\infty)$

  - Algorithm: Composition

    – $Y \sim N(\mu,\sigma^2) \Rightarrow X = e^Y$

  - Expectation: $E(X) = e^{\mu+\frac{\sigma^2}{2}}$

  - Variance: $VAR(X) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$

Note that μ and σ are NOT the mean and the variance of the lognormal distribution!

Distribution - Continuous

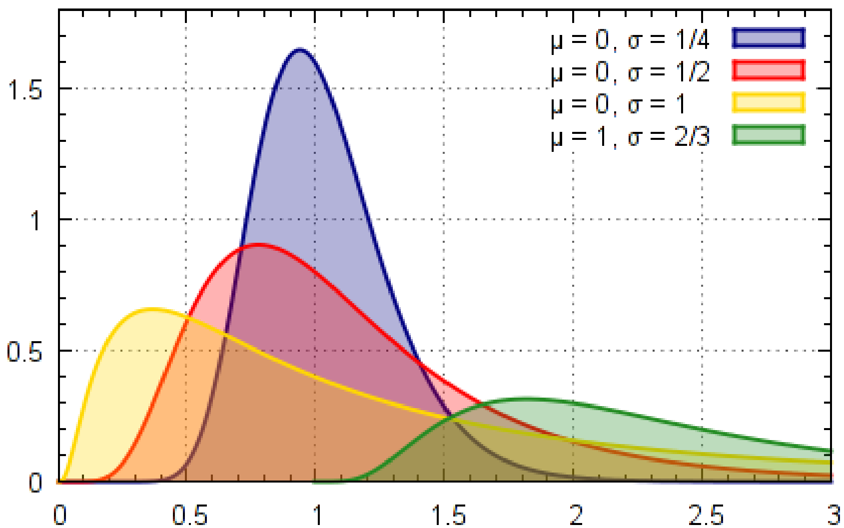
- **Lognormal distribution(2/3):** $RV\ \ X \sim LN(\mu,\sigma^2)$ (LK 8.3.7)

  - Parameters of the normal distribution which is used to generate LN

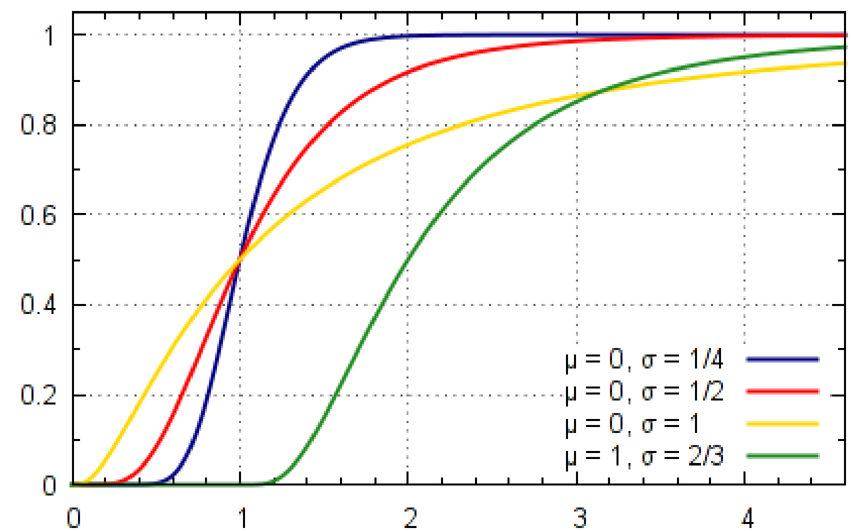
    – $$\mu = E[Y] = \ln\left(\frac{E[X]^2}{\sqrt{E[X]^2 + VAR[X]}}\right)$$

    – $$\sigma^2 = VAR[Y] = \ln\left(\frac{E[X]^2}{\sqrt{E[X]^2 + VAR[X]}}\right)$$

- **Lognormal distribution(3/3):** $RV\ \ X \sim LN(\mu,\sigma^2)$ (LK 8.3.7)
  Use case: risk management (insurance companies)



Probability Density Function



Cumulative Density Function

Distribution - Continuous

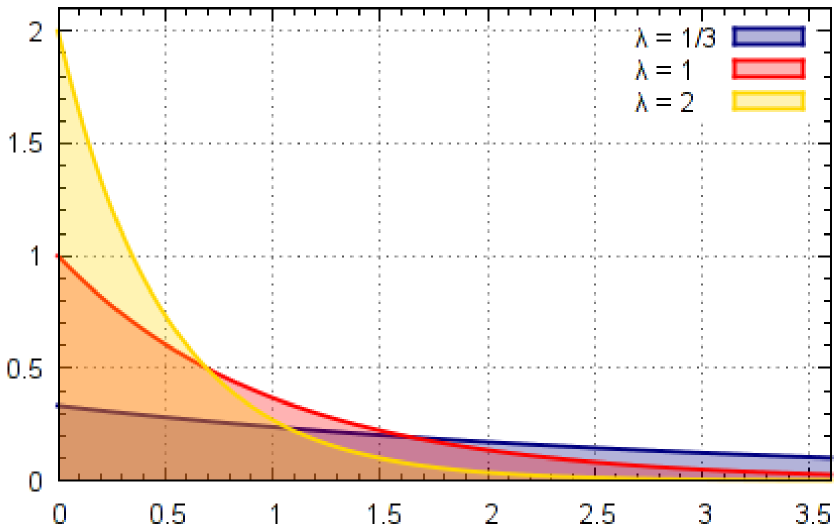
- **Exponential distribution(1/2):** $RV \;\; X \sim \exp(\lambda)$ (LK 8.3.2)

  - Density function: $f(x) = \lambda \cdot e^{-\lambda x} \quad \text{für} \quad x \geq 0$
  - Distribution function: $F(x) = 1 - e^{-\lambda x}$
  - Range: $[0, \infty[$ Mode: 0
  - Expectation: $E(X) = \frac{1}{\lambda}$
  - Variance: $VAR(X) = \frac{1}{\lambda^2}$
  - Coefficient of variation: $c_{Var} = 1$
  - Generation: Inversion $U \sim U(0,1), X = \frac{-\ln(U)}{\lambda}$

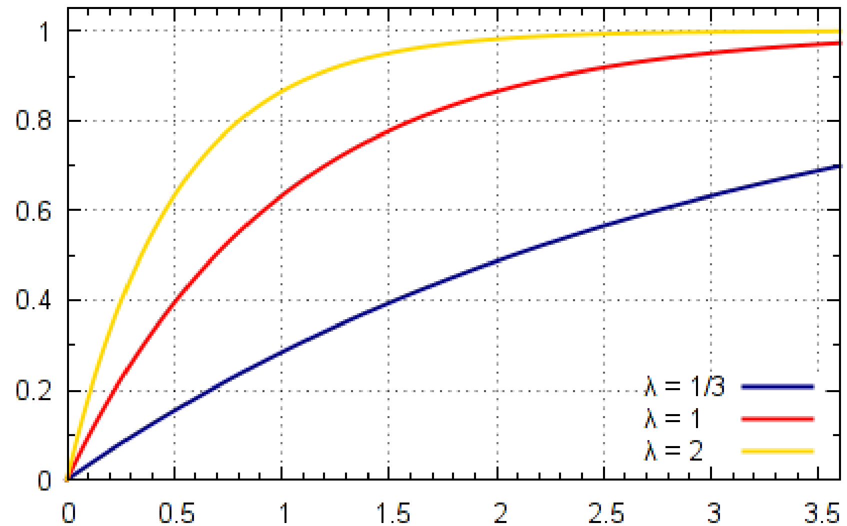
Distribution - Continuous

- **Exponential distribution(2/2):** $RV \ \ X \sim \exp(\lambda)$ (LK 8.3.2)

  Use case: life time of structures, time between calls/requests



Probability Density Function



Cumulative Density Function

Pictures taken from Wikipedia

Distribution - Continuous

- **Erlang-k distribution(1/3):** $RV \;\; X \sim k - Erlang(\lambda)$ (LK 8.3.3)

  - $RV \;\; X = Y_1 + Y_2 + Y_3 + \cdots + Y_k$ where the Yi's are IID exponential random variables

  - Density function:

$$f(x) = \begin{cases} \dfrac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} & for \quad x \geq 0 \\ 0 & Otherwise \end{cases}$$

  - Distribution function:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} \cdot \sum_{i=0}^{k-1} \dfrac{(\lambda x)^i}{i!} & for \quad x \geq 0 \\ 0 & Otherwise \end{cases}$$

RV X represents the sum of k exponential random variables

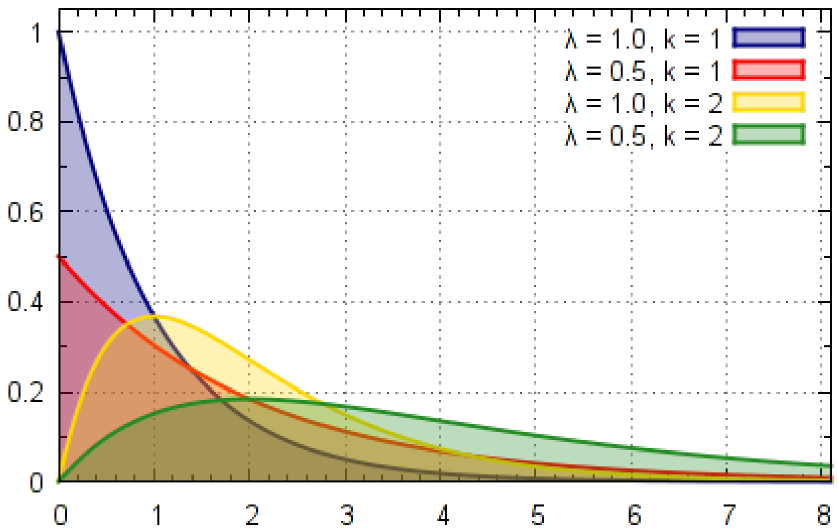
Distribution - Continuous

- **Erlang-k distribution(2/3):** $RV\ \ X \sim k - Erlang(\lambda)$ (LK 8.3.3)
  - Range: $[0, \infty[$
  - Expectation: $E(X) = \frac{k}{\lambda}$
  - Variance: $VAR(X) = \frac{k}{\lambda^2}$
  - Mode: $\frac{k-1}{\lambda}$
  - Coefficient of variation: $c_{Var} = \frac{1}{\sqrt{k}}$
  - Generation:
    - » Inversion $U_i \sim U(0,1), X = \frac{-\ln\left(\prod_{0 \le i < k} U_i\right)}{\lambda}$
    - » Convolution $RV\ \ X = Y_1 + Y_2 + Y_3 + \cdots + Y_k$

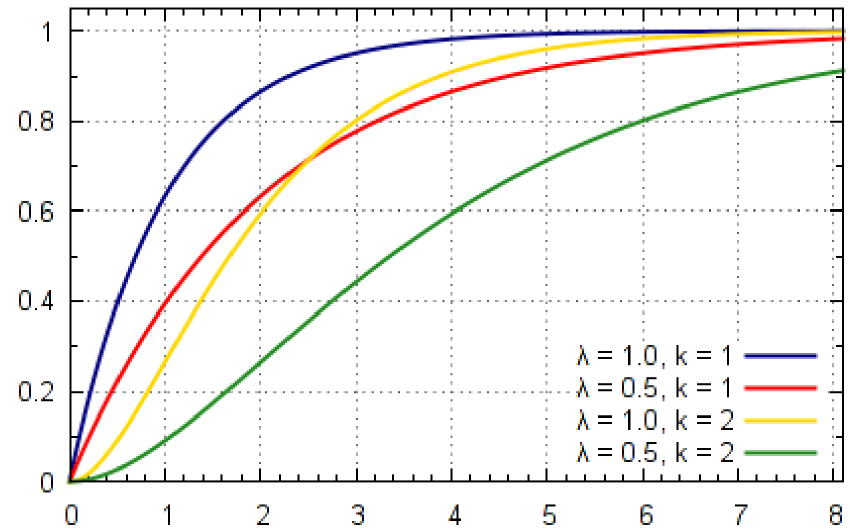
Distribution - Continuous

- **Erlang-k distribution(3/3):** $RV \;\; X \sim k - Erlang(\lambda)$ (LK 8.3.3)

  Use case: lifetime of structures, delay in transport networks, dimensioning of systems (e.g. call center)



Probability Density Function



Cumulative Density Function

Distribution - Continuous

- **Gamma distribution(1/3):** $RV \;\; X \sim gamma(\alpha, \beta)$ (LK 8.3.4)
  - Density function: $$f(x) = \begin{cases} \dfrac{\beta^{-\alpha} \cdot x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)} & for \quad x \leq 0 \\ 0 & Otherwise \end{cases}$$
  - Distribution function: $$F(x) = \begin{cases} 1 - e^{\frac{-x}{\beta}} \cdot \sum\limits_{0 \leq i < \alpha} \dfrac{\left(\dfrac{-x}{\beta}\right)^i}{i!} & for \quad x > 0 \\ 0 & Otherwise \end{cases}$$
  - Parameter description:
    - Location parameter γ: Shifting the distribution along the x-axis
    - Scale parameter β: Linear impact on the expectation
    - Shape parameter α: Changes the shape of the distribution

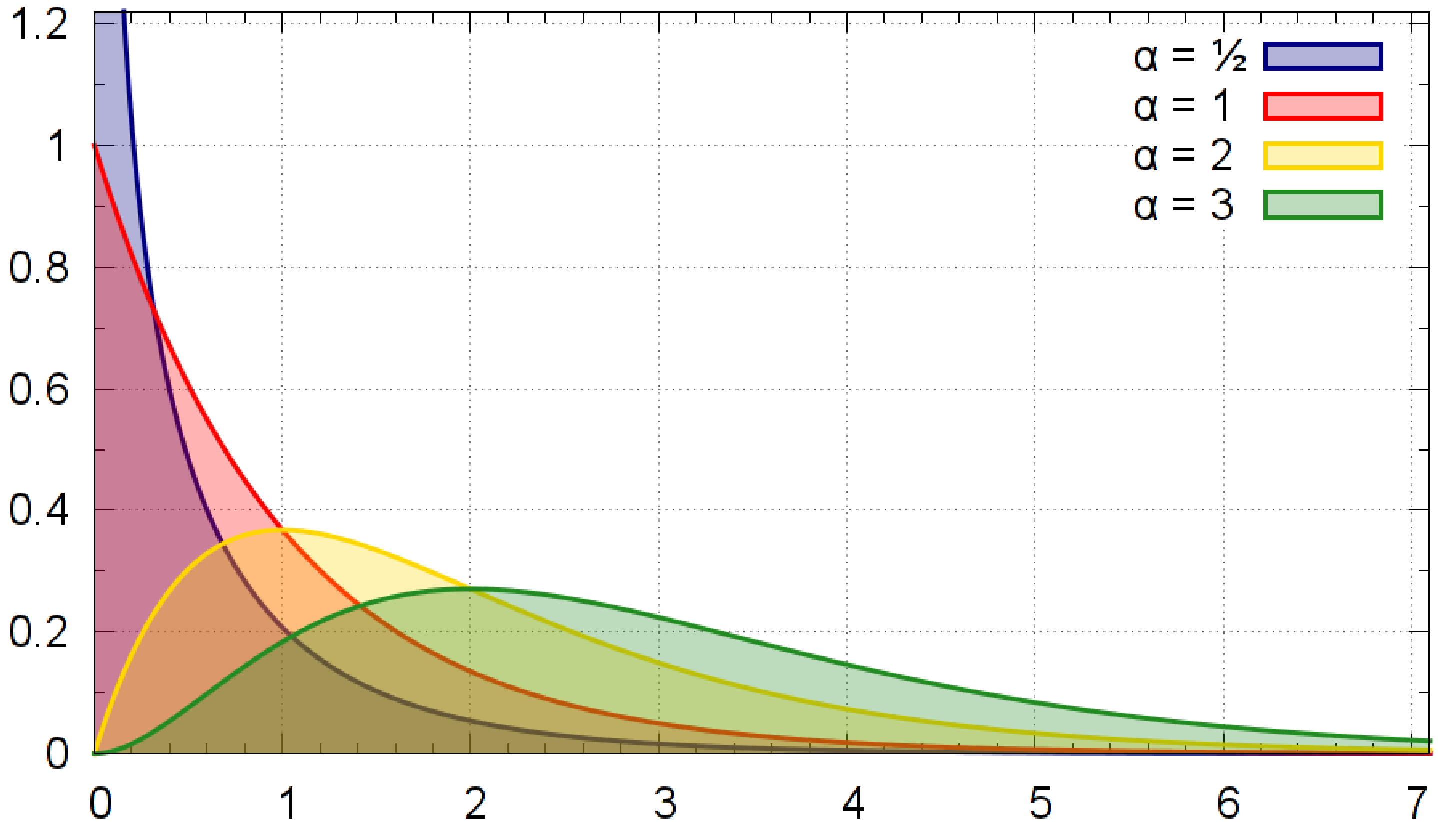
Distribution - Continuous

- **Gamma distribution(2/3):** $RV\ \ X \sim gamma(\alpha, \beta)$ (LK 8.3.4)
  - Gamma function: $$\Gamma(z) = \begin{cases} \int_0^\infty t^{z-1} e^{-t} dt & if \ \ x \geq 0 \\ 0 & if \ \ x < 0 \end{cases}$$
  - Expectation: $E(X) = \alpha \cdot \beta$
  - Coefficient of variation: $c_{Var} = 1$
  - Mode: $$\begin{cases} 0 & if \ \ \alpha < 1 \\ \beta \cdot (\alpha - 1) & if \ \ \alpha \geq 1 \end{cases}$$
  - Generation:
    - Step 1 $X \sim gamma(\alpha, \beta) \rightarrow X = \beta \cdot Y \qquad Y \sim gamma(\alpha, 1)$
    - Step 2 Generation of $X \sim gamma(\alpha, 1)$ with Accept-Reject

Distribution - Continuous

- **Gamma distribution(3/3):** $RV \ \ X \sim gamma(\alpha, \beta)$ (LK 8.3.4)

  Use cases: risk management (insurance companies), service time, down time



Probability Density Function

Distribution - Discrete

- **Uniform (discrete) (1/2)** $RV\ \ X \sim DU(i, j)$ (LK 8.4.2)

  - Distribution: $$p(k) = \begin{cases} \dfrac{1}{j-i+1} & if\ \ k \in \{i, i+1, i+2, \ldots, j\} \\ 0 & Otherwise \end{cases}$$
  - Range: $i \le k \le j$
  - Expectation: $$E(X) = \frac{(i+j)}{2}$$
  - Variance: $$VAR(X) = \frac{(j-i+1)^2 - 1}{12}$$
  - Generation: Inversion

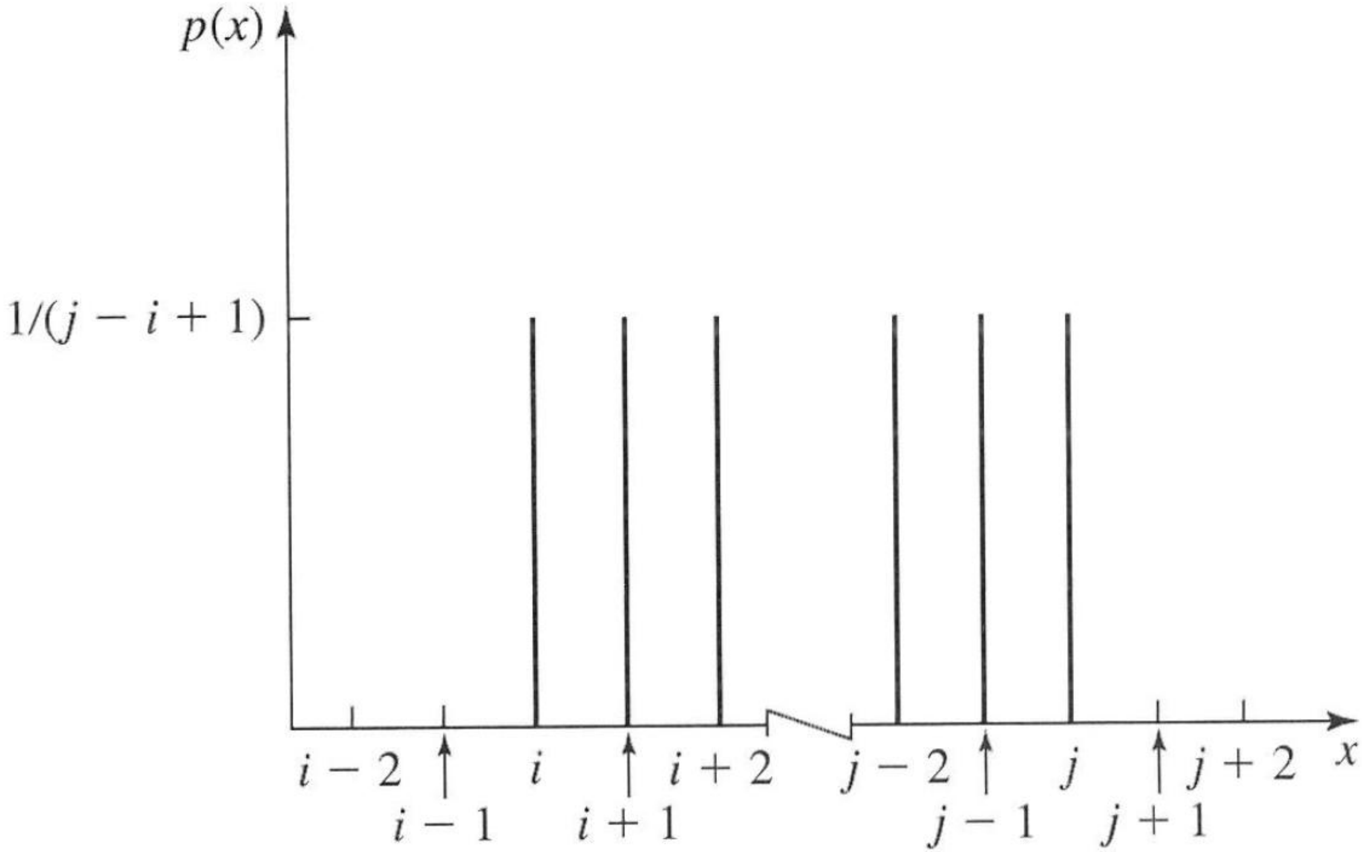
    $U \sim U(0,1) \qquad X = i + \lfloor (j-i+1) \cdot U \rfloor$

DU(0,1) and Bernoulli(0.5) distributions are the same

Distribution - Discrete

- **Uniform (discrete) (2/2)** $RV\ \ X \sim DU(i, j)$ (LK 8.4.2)

  Use case: backoff distribution, simulation (dice, roulette, …)



Distribution

Distribution - Discrete

- **Bernoulli (1/2)** $RV\ \ X \sim Bernoulli\ (p)$ (LK 8.4.1)
  - Example: Flipping a coin 
  - Distribution: $$p(k) = \begin{cases} 1-p & if\ \ k=0 \\ p & if\ \ k=1 \\ 0 & Otherwise \end{cases}$$
  - Range: $$i \le k \le j$$
  - Expectation: $$E(X) = p$$
  - Variance: $$VAR(X) = p \cdot (1-p)$$
  - Coefficient of variation: $$c_{Var} = \sqrt{\frac{1-p}{n \cdot p}}$$

Distribution - Discrete

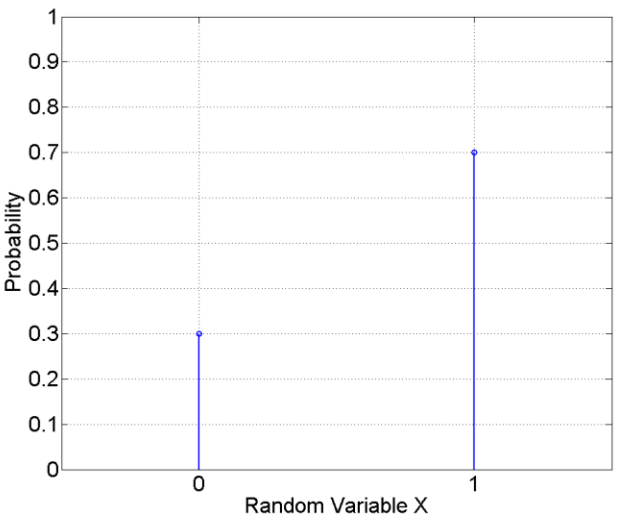
- **Bernoulli (2/2)** $RV\ \ X \sim Bernoulli\ \ (p)$ (LK 8.4.1)

  - Mode: 0 or 1 (depends on the definition of the outcome)
  - Generation: Inversion $U \sim U(0,1)$

$$X = \begin{cases} 0 \ \ if \ \ U < p \\ 1 \ \ Otherwise \end{cases}$$

  - Distribution

    $Bernoulli\ \ (0.3)$



Distribution - Discrete

- **N-Bernoulli (1/2)** $RV \;\; X \sim Bernoulli\;(n, p)$ (LK 8.4.4)

  - Example: Flipping a coin n times

  - Distribution: $p(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \qquad 0 \le k \le n$
  - Range: $0 \le k \le n$
  - Expectation: $E(X) = np$
  - Variance: $VAR(X) = n \cdot p \cdot (1-p)$
  - Coefficient of variation: $c_{Var} = \sqrt{\dfrac{1-p}{n \cdot p}}$
  - Use case: quality management, wrong/right decisions

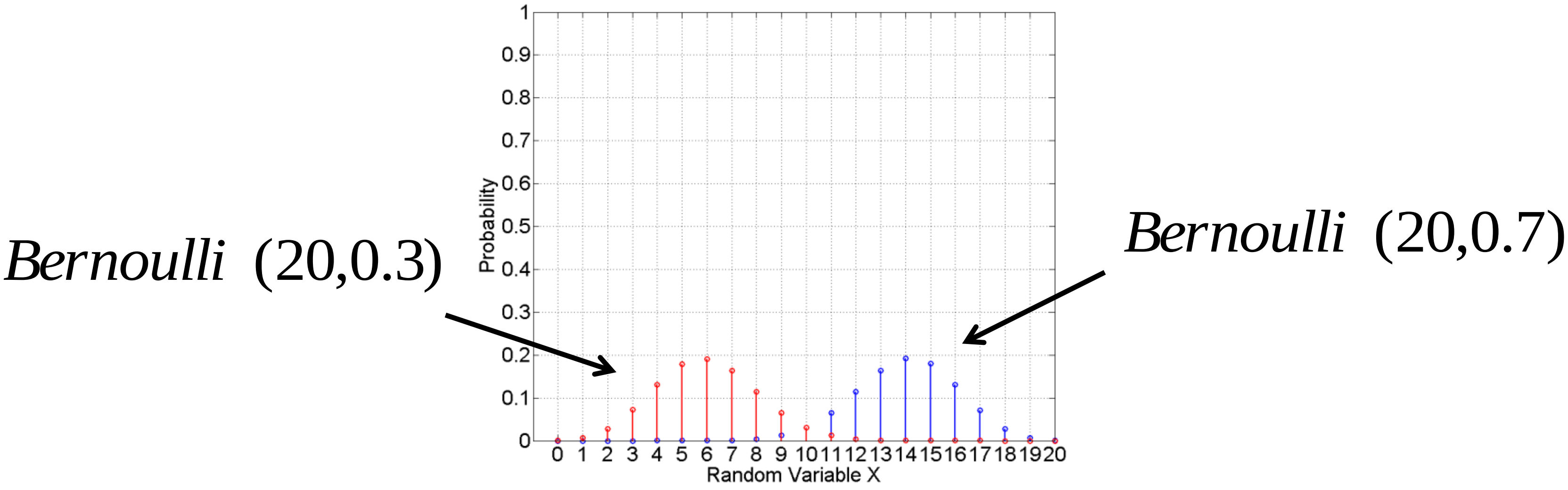
Distribution - Discrete

- **N-Bernoulli (2/2)** $RV \;\; X \sim Bernoulli \;(n, p)$ (LK 8.4.4)

  - Mode: 0 or 1 (depends on the definition of the outcome)
  - Generation: Composition

$$Bernoulli \;(n, p) \approx \sum_{0 \le i < n} Bernoulli \;(p)$$

  - Distribution



Distribution - Discrete

- **Geom (1/2)** $RV\ \ X \sim Geom\,(p)$ (LK 8.4.5)

  - Example: Number of unsuccessful Bernoulli – Experiments **until** a successful outcome (e.g. number of retransmissions)

  - Distribution: $p(x) = p \cdot (1-p)^x$

  - Distribution function: $F(x) = 1-(1-p)^{\lfloor x \rfloor + 1}$

  - Expectation: $E(X) = \dfrac{1-p}{p}$

  - Variance: $VAR(X) = \dfrac{1-p}{p^2}$

  - Coefficient of variation: $c_{Var} = \sqrt{\dfrac{1}{1-p}}$

Distribution - Discrete

- **Geom (2/2)** $RV \quad X \sim Geom\,(p)$ (LK 8.4.5)

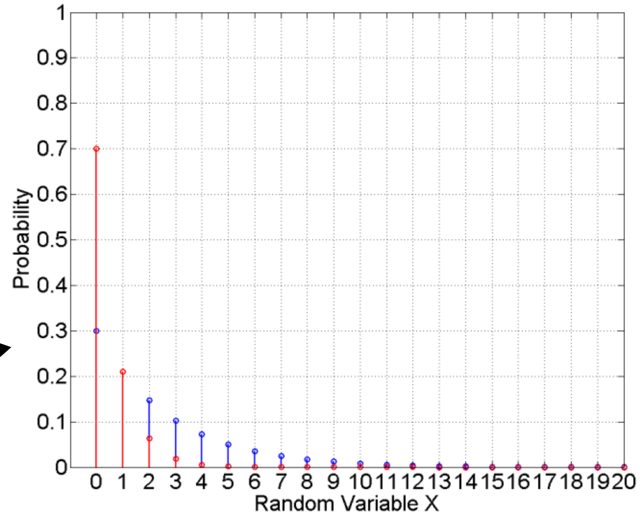
  - Mode: 0
  - Generation: Inversion $U \sim U(0,1)$

$$X = \left\lfloor \frac{\ln(U)}{\ln(1-p)} \right\rfloor$$

  - Use case: delivery ratio in computer networks, risk management
  - Distribution

$Geom\,(0.7) \rightarrow$

$Geom\,(0.3) \rightarrow$



$$p(0) = p$$

Distribution - Discrete

- **Poisson(1/3)** $RV \;\; X \sim Poisson\ (\lambda)$ (LK 6.2.4)

  - Example: Number of events that occur in an interval of time when the events are occurring at a constant rate (number of items in a batch of random size)
  - Distribution: $$p(x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda} \qquad \text{if } x \in \{0,1,2,\ldots\}$$
  - Distribution function: $$F(x) = \begin{cases} e^{-\lambda} \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$
  - Parameter: $\lambda > 0$

Distribution - Discrete

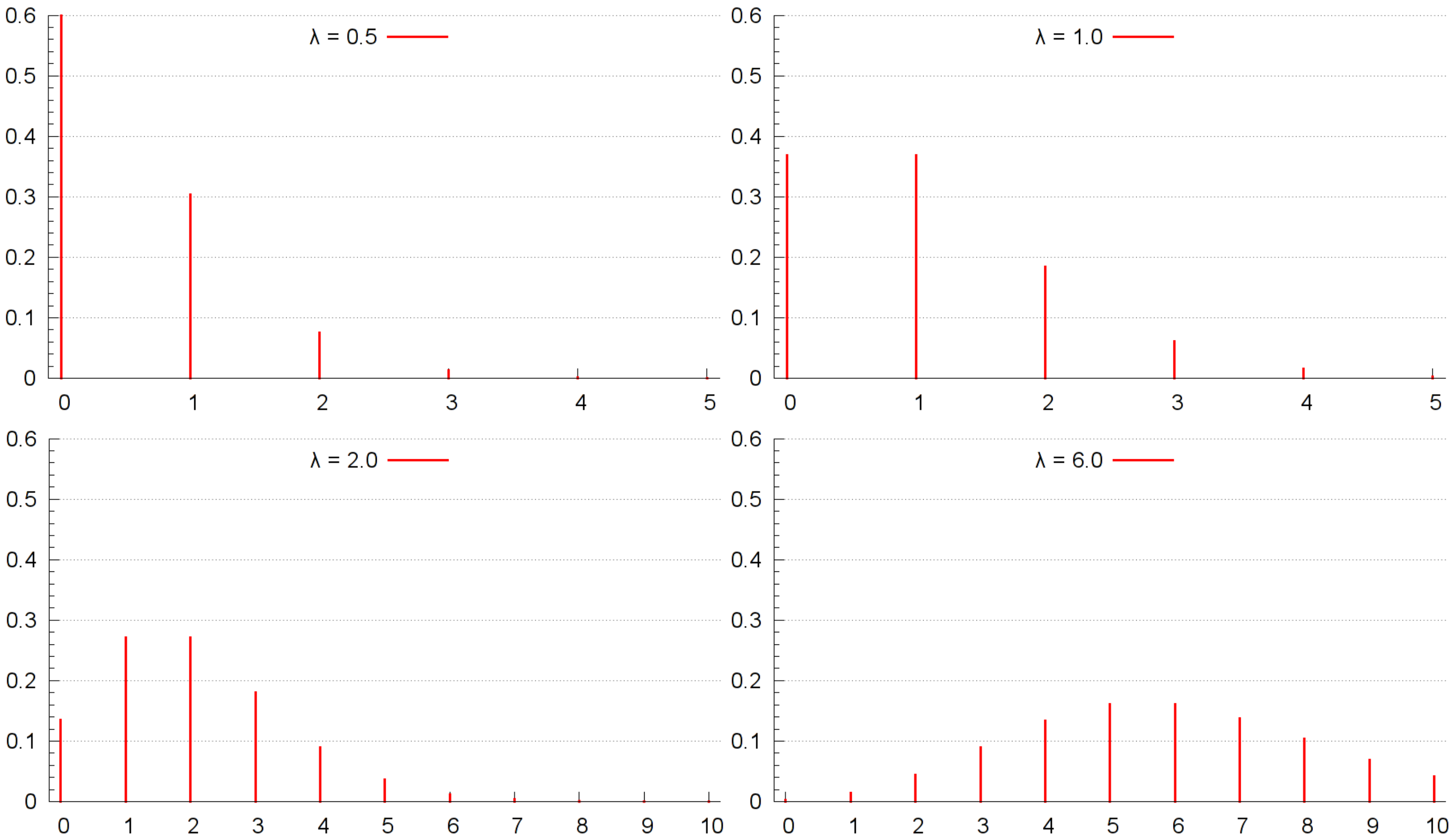
Poisson(2/3) $RV \;\; X \sim Poisson\;(\lambda)$ (LK 6.2.4)

- Range: $\{0,1,2,3,\ldots\}$
- Expectation: $E(X) = \lambda$
- Variance: $VAR(X) = \lambda$
- Coefficient of variation: $c_{Var} = \dfrac{1}{\sqrt{\lambda}}$
- Mode $\begin{cases} \lambda \cap \lambda - 1 & \lambda \text{ is an integer} \\ \lfloor \lambda \rfloor & \text{otherwise} \end{cases}$
- Special characteristics:
  - $x = 0$ ⇨ exponential distribution (time interval between two consecutive events)
  - Number of events until a certain point in time is Poisson distributed
  - Period of time until n events have occurred is Erlang distributed

Distribution - Discrete

- **Poisson(3/3)** $RV \;\; X \sim Poisson\ (\lambda)$ (LK 6.2.4)

  Use case: number of (independent) arrivals in a certain time interval



Distribution - Discrete

- **General Discrete(1/1)** $RV \;\; X \sim GD$ (LK 8.4.3)

  - Distribution: $$p(x) = \begin{cases} p_k & if \;\; x = x_k, \;\; 0 \le k < n \\ 0 & Otherwise \end{cases}$$

  - Generation: Inversion $U \sim U(0,1)$

$$X = x_k \text{ , falls } \sum_{j=0}^{k-1} p_j \le U < \sum_{j=0}^{k} p_j$$