# Review (2)

Area de Ingeniería Telemática
http://www.tlm.unavarra.es

Máster en Tecnologías Informáticas

# Contents

- Probability review and tips
  - Random variables
  - Random number generation
  - Basic modeling
  - Poisson process

# ¿ Why random variables ?

- Imagine the time it takes a user to download a Web resource
- **It depends on:** The size of the resource, how fast the web server disk is, the load the disk is serving, how powerful the CPU of the server is, how fast the server bus is, how many other devices are using that bus, how many other processes are using the CPU and how, how much RAM/L1-3cache the server has and whether it is paging/swapping, how the web server writes in the TCP buffer (size of the chunks), the flow control TCP buffer size in the client, the buffer size used by the TCP server, how much traffic (and how) is the server sending/receiving through the NIC, the network between client and server (delay, loss or not for each packet), the Path MTU, the timer values configured in the server and client (delayed ACK, retransmission timers), the power of the client CPU, the implementation of TCP in the client, how the client retrieves the data from the TCP buffer, the RAM size at the client, how many other processes are running in the client, etc etc etc…
- Too many parameters !!!!
- It is much easier to describe the world in a probabilistic way than in a deterministic one

# Probability

- A **random variable** (r.v.) X is the outcome of a random event expressed as a numeric value
- The *Cummulative Distribution Function (CDF)* provides the fixed probability that the r.v. will not exceed a value *x*

$$CDF(X) \equiv F_X(x) \equiv P(X \leq x)$$

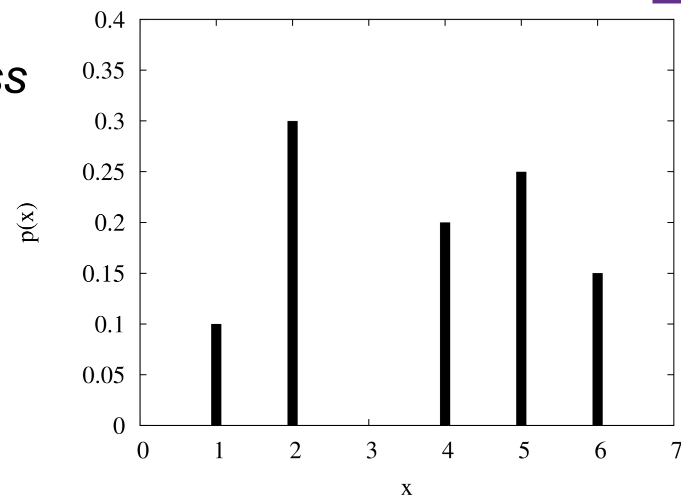
- The *Complementary Cummulative Distribution Function (CCDF)*:

$$CCDF(X) \equiv \overline{F}_X(x) \equiv 1 - F_X(x) \equiv P(X > x)$$

- **Discrete** r.v. : takes values from a finite or a countably infinite set of values
- *Probability Distribution* or *Probability Mass Function* of a discrete r.v. :

$$p_X[x_i] \equiv P(X = x_i)$$

$$p_X\left[x_i\right] \geq 0 \qquad \sum_{i=1}^{\infty} p_X\left[x_i\right] = 1$$



# Continuous rr.vv.

- **Continuous** r.v. : takes values from an uncountably infinite set of values $R_X$
- *Probability Density Function* of a continuous r.v. :

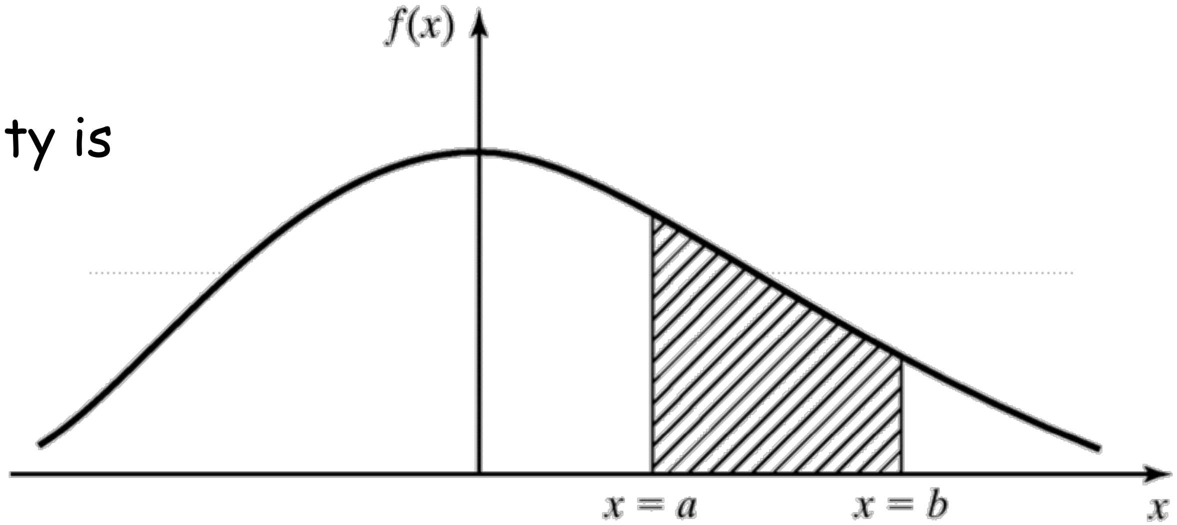
$$f_X(x) \equiv \frac{dF_X(x)}{dx} = \frac{dP(X \le x)}{dx}$$

$$P(x_1 < X \le x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(u)du$$

$$f_X(x) \ge 0 \quad (x \in R_X) \qquad \int_{R_X} f_X(x)dx = 1$$

$$P(x_1 < X \le x_2) = P(x_1 \le X \le x_2) = P(x_1 \le X < x_2) = P(x_1 < X < x_2)$$

The probability is in the area



# Moments

- *Expected value* of a continuous random variable X (a.k.a. expectation, mean, first moment):

$$E[X] \equiv \mu_X \equiv \int_{-\infty}^{\infty} u\, f_X(u)du$$

- *n*th *moment* of X: $E[X^n] \equiv \int_{-\infty}^{\infty} u^n f_X(u)du$

- Related with the variability is the *variance* :

$$Var(X) \equiv \sigma_X{}^2 = E\left[(X-\mu_X)^2\right] = \int_{-\infty}^{\infty}(u-\mu_u)^2 p(u)du = E\left[X^2\right]-\left(E[X]\right)^2 = E\left[X^2\right]-\mu_X{}^2$$

- *Standard deviation*: $\sigma_X \equiv \sqrt{Var(X)}$

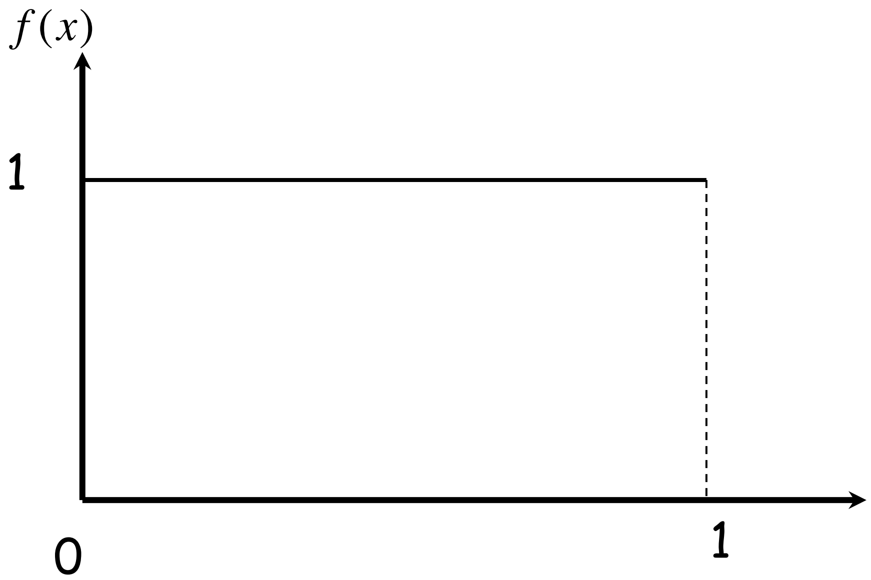
- *Coefficient of variation:* $c_v = \dfrac{\sigma_X}{\mu_X}$

# Commonly Encountered Distributions

| Distribution | Definition | Domain |
|---|---|---|
| Exponential | $p(x) = \lambda e^{-\lambda x}$ | $x > 0$ |
| Normal | $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$ | $-\infty < x < \infty$ |
| Gamma | $p(x) = \frac{(x-\gamma)^{\alpha-1} \exp[-(x-\gamma)/\beta]}{\beta^\alpha \Gamma(\alpha)}$ | $x > \gamma$ |
| Extreme | $F(x) = \exp\left[-\exp\left(-\frac{(x-\alpha)}{\beta}\right)\right]$ | $-\infty < x < \infty$ |
| Lognormal | $p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2\right]$ | $x > 0$ |
| Pareto | $p(x) = \alpha k^\alpha x^{-\alpha-1}$ | $x > k$ |
| Weibull | $p(x) = \frac{bx^{b-1}}{a^b} \exp\left[-\left(\frac{x}{a}\right)^b\right]$ | $x > 0$ |

# Random number generation

- We first try to generate random numbers from a uniform distribution
- Independent



# Pseudo-random numbers

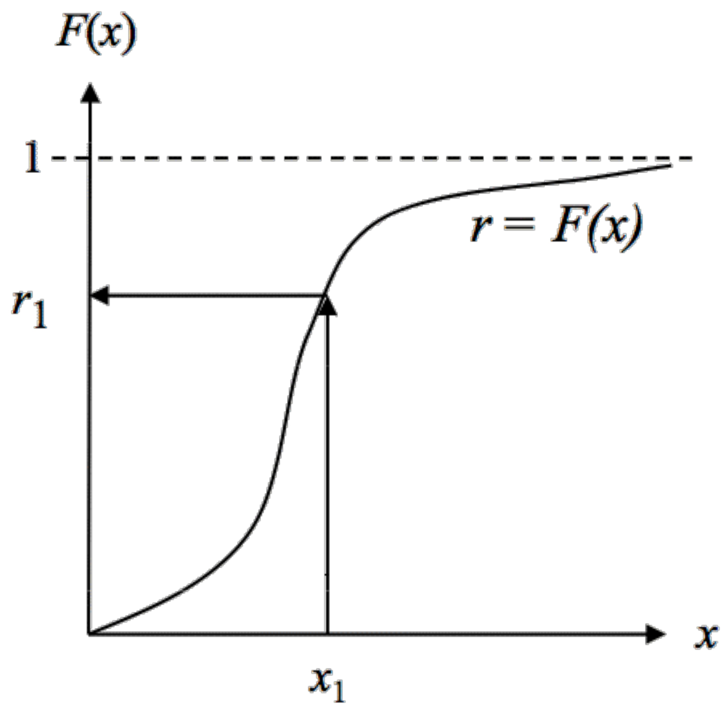
- They look like random
- Known the seed they are predictable
- They even have a period
- Example: Linear Congruential Method

$$X_{i+1} = \left(aX_i + c\right) \bmod m$$

- What about a non uniform distribution?

# Inverse-transform Technique

- F(x) is the CDF of the target r.v.
- X uniform r.v. in [0,1]
- Generate a sample $r_1$ from X
- Use the inverse function to obtain $x_1 = F^{-1}(x)$
- $x_1$ is a sample from a r.v. with CDF F(x)
- Of course it is easier if F(x) has a simple analytical inverse



# Example: Exponential distribution

$$f(x) = \lambda e^{-\lambda x}$$

$$F(x) = 1 - e^{-\lambda x}$$

$$R = 1 - e^{-\lambda X}$$

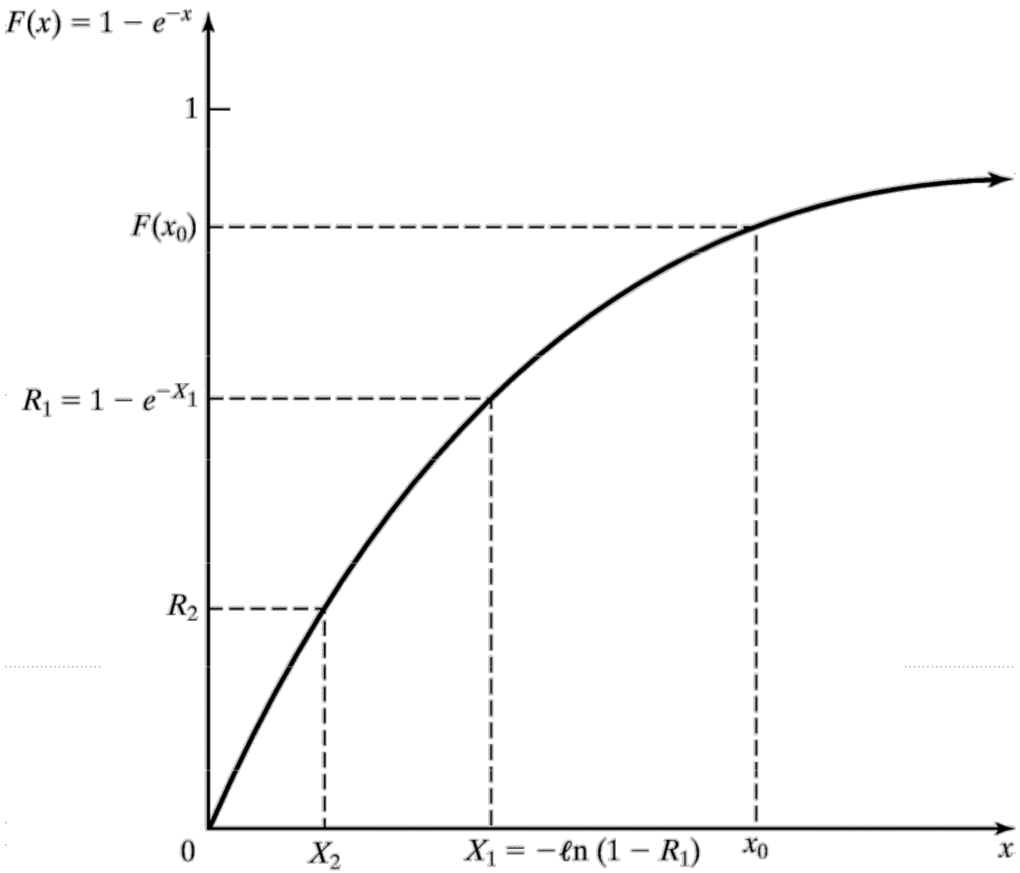
$$1 - R = e^{-\lambda X}$$

$$\ln(1 - R) = -\lambda X$$

$$X = -\frac{\ln(1-R)}{\lambda} = F^{-1}(R)$$

ó

$$X = -\frac{\ln(R)}{\lambda} = F^{-1}(R)$$




(Both R and 1-R are uniform rr.vv.)

# Inverse-transform Technique

- "Easy" distributions: Triangular, Weibull, Pareto
- F(x) could come from experimental samples
  - Use interpolation for a little improvement
- For discrete rr.vv. only a table is needed
- "Hard" ones: Gamma, Normal, Beta
- Numerical approximations to the CDF or to the inverse CDF could also be useful

# Techniques based on properties

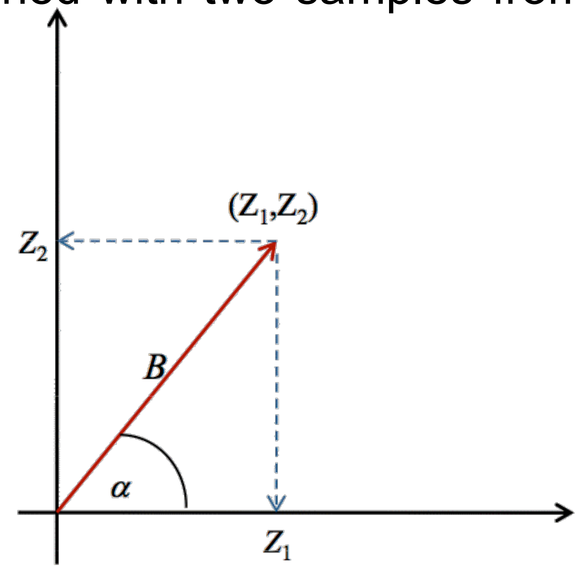
**Example: Gaussian distribution**

- $Z_1$ and $Z_2$ rr.vv. ϕ(0,1)
- They are the rectangular coordinates of a point ($Z_1$,$Z_2$)
- In polar coordinates: $\begin{cases} Z_1 = B\cos(\alpha) \\ Z_2 = B\sin(\alpha) \end{cases}$
- The radial coordinate B is a r.v. from an exponential distribution
- The angular coordinate is a r.v. from an uniform distribution
- They are independent
- So two samples from ϕ(0,1) can be obtained with two samples from an uniform distribution

$$Z_1 = \sqrt{-2\ln(R_1)}\cos(2\pi R_2)$$
$$Z_2 = \sqrt{-2\ln(R_1)}\sin(2\pi R_2)$$

- And from Y = ϕ(μ,σ) :

$$Y = \mu + \sigma Z_i$$



# Building a model

- Sample the phenomenon
- Select a known distribution that “is similar”
- Estimate the parameters of this distribution
- Test to see how good the fit is for the original purpose

# Building a model: example

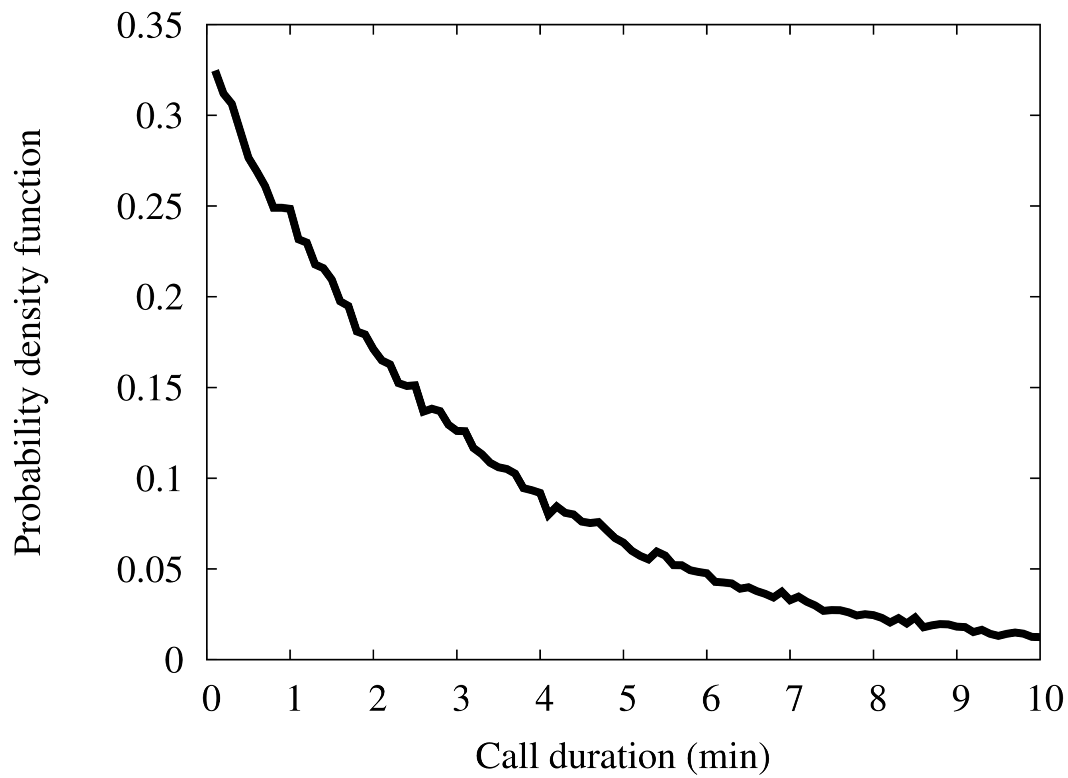
- Sample the phenomenon
  - Duration of phone calls
- Select a known distribution that "is similar"
- Estimate the parameters of this distribution
- Test to see how good the fit is for the original purpose

| Call durations (minutes) |
|---|
| 8.2947495235 |
| 2.1268147168 |
| 0.5884509608 |
| 3.5020706914 |
| 5.2125237671 |
| 2.8848404480 |
| 6.2123475174 |
| 4.2605010872 |
| … |

# Building a model: example

- Sample the phenomenon
- Select a known distribution that "is similar"
  - Example: visual inspection… mmm… looks like exponential
- Estimate the parameters of this distribution
- Test to see how good the fit is for the original purpose

| Call durations (minutes) |
|---|
| 8.2947495235 |
| 2.1268147168 |
| 0.5884509608 |
| 3.5020706914 |
| 5.2125237671 |
| 2.8848404480 |
| 6.2123475174 |
| 4.2605010872 |
| … |

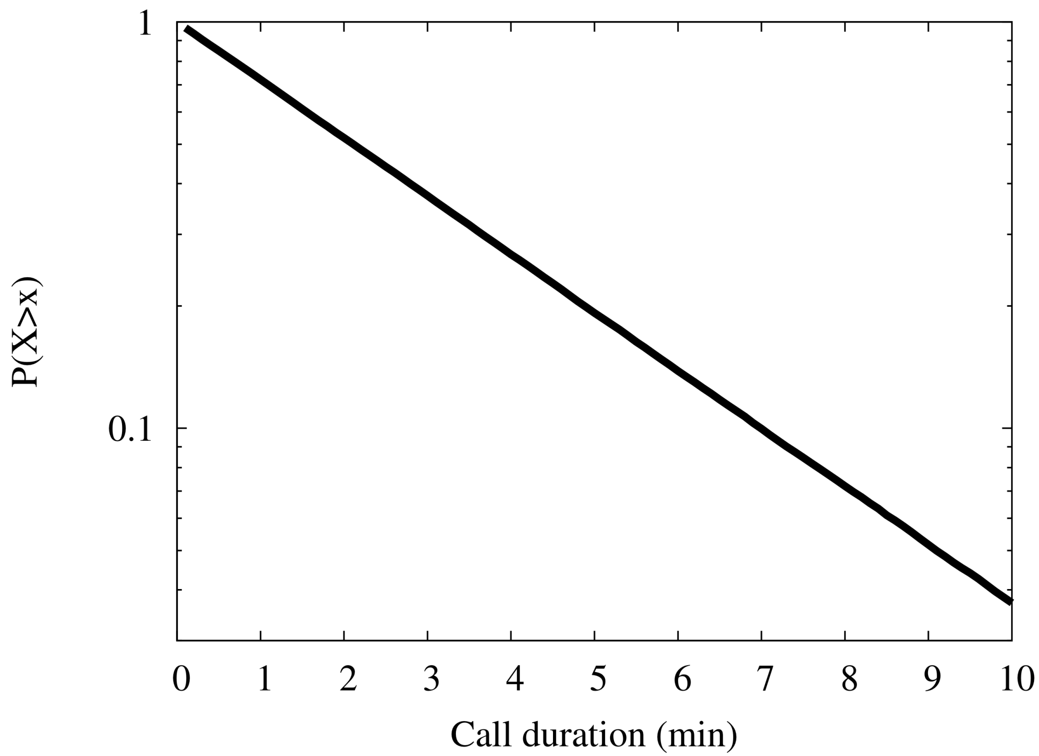


# Building a model: example

- Sample the phenomenon
- Select a known distribution that "is similar"
- Estimate the parameters of this distribution
  - Example: for exponential distribution, CCDF in a log-linear plot
  - Use least squares fitting to estimate the slope
- Test to see how good the fit is for the original purpose

$$P[X_i > t] = e^{-\lambda t}$$

| Call durations (minutes) |
|---|
| 8.2947495235 |
| 2.1268147168 |
| 0.5884509608 |
| 3.5020706914 |
| 5.2125237671 |
| 2.8848404480 |
| 6.2123475174 |
| 4.2605010872 |
| … |

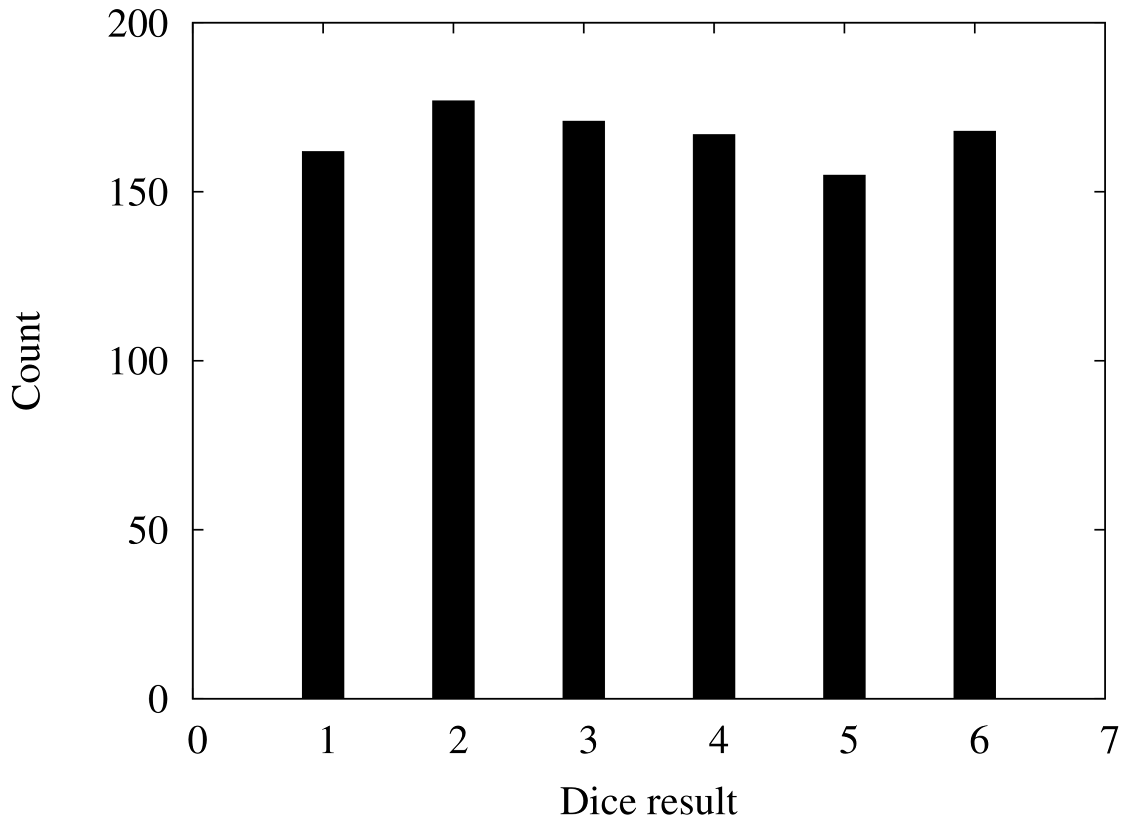


# Drawing a distribution

## Discrete r.v.

- Obtain samples
- Compute histogram

| Dice result | count |
|---|---|
| 1 | 162 |
| 2 | 177 |
| 3 | 171 |
| 4 | 167 |
| 5 | 155 |
| 6 | 168 |

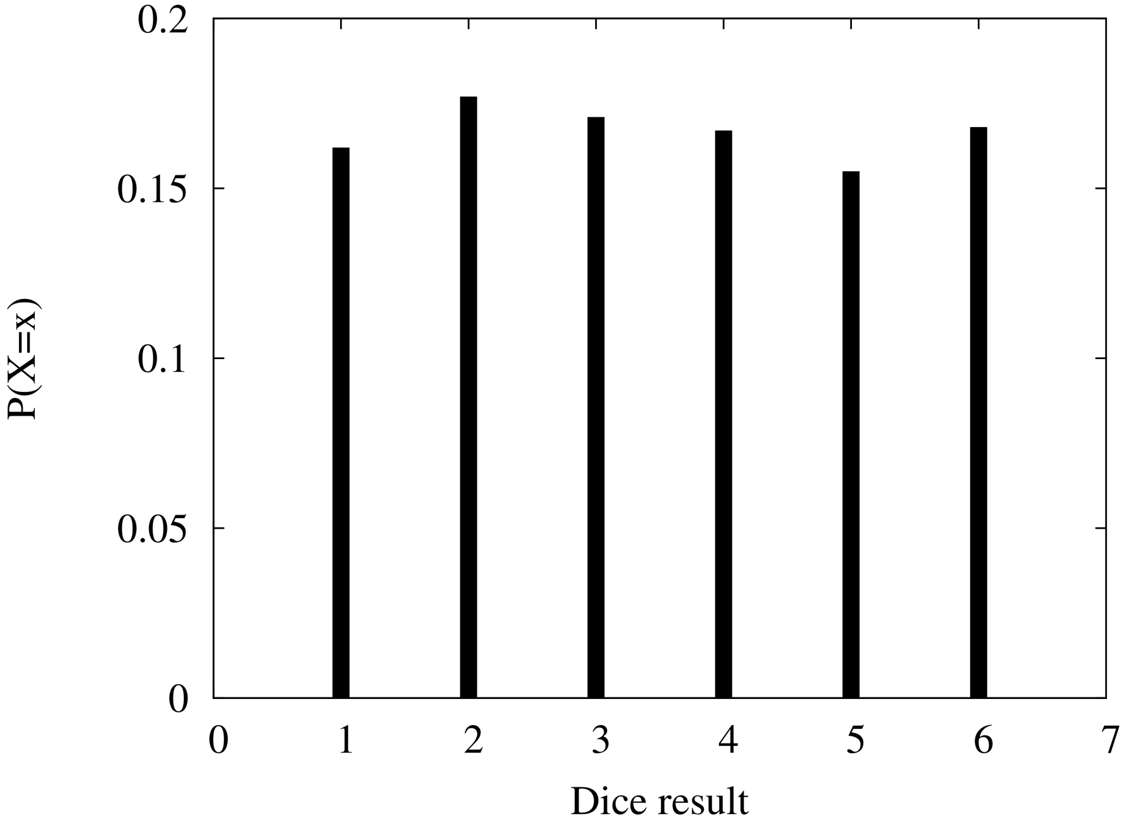


# Drawing a distribution

## Discrete r.v.

- Obtain samples
- Compute histogram
- Estimate probabilities based on occurrence count (divide by total number of samples)

| Dice result | count |
|---|---|
| 1 | 162 |
| 2 | 177 |
| 3 | 171 |
| 4 | 167 |
| 5 | 155 |
| 6 | 168 |

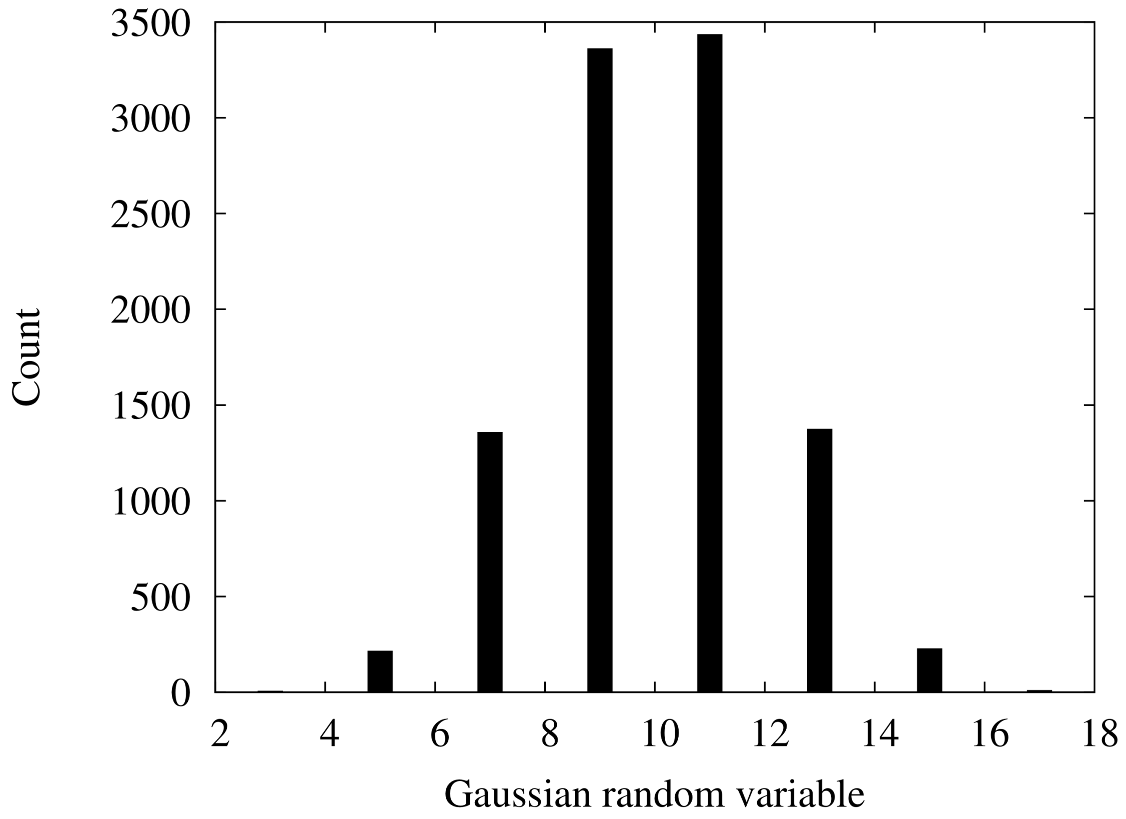


# Drawing a distribution

## Continuous r.v.

- Obtain samples
- Compute a histogram (decide boxes width)

| Cube | count |
|---|---|
| [2,4) | 8 |
| [4,6) | 217 |
| [6,8) | 1359 |
| [8,10) | 3363 |
| [10,12) | 3437 |
| [12,14) | 1376 |
| [14,16) | 229 |
| [16,18) | 11 |

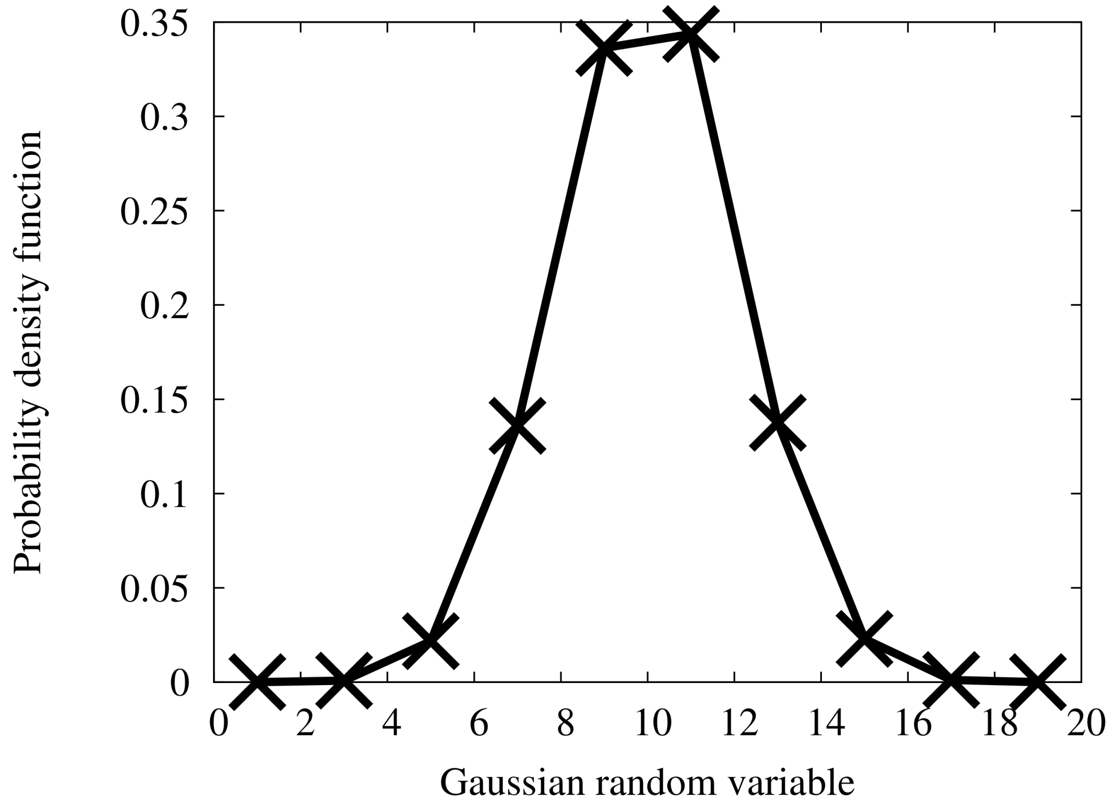


# Drawing a distribution

## Continuous r.v.

- Obtain samples
- Compute a histogram (decide boxes width)
- Estimate probabilities based on occurrence count (divide by total number of samples). Is that all?

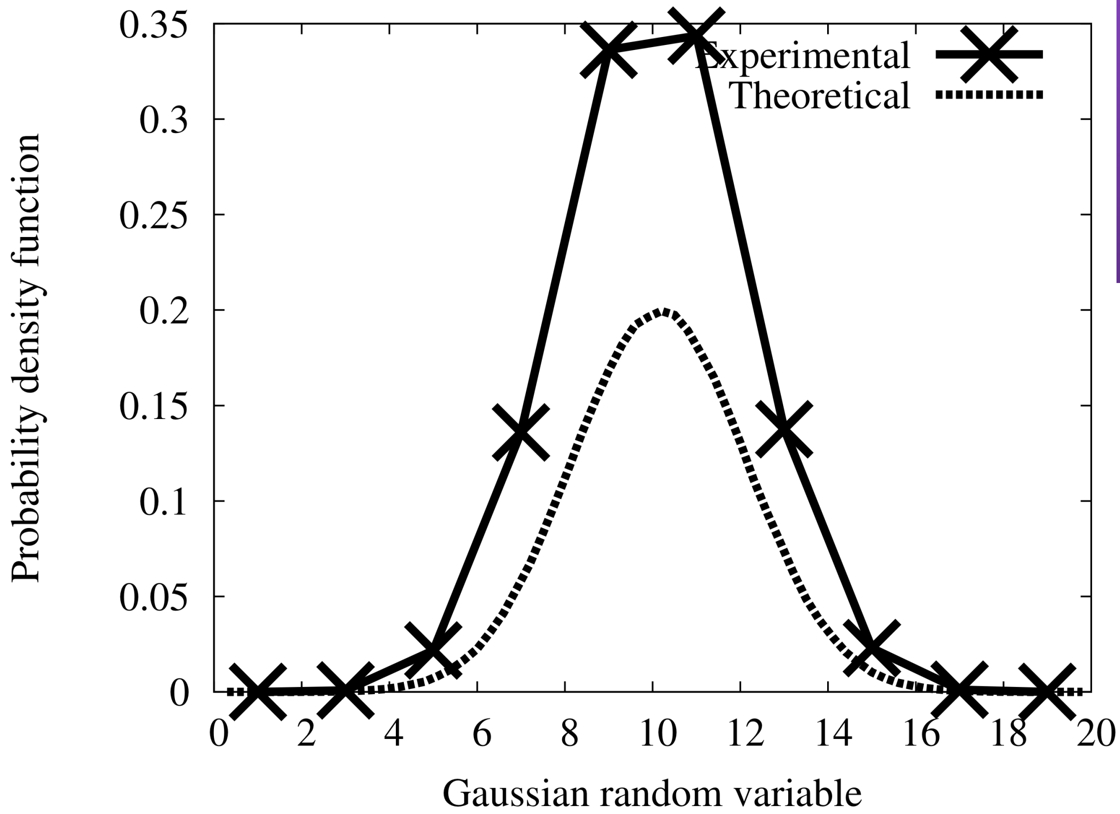
| Cube | count |
|---|---|
| [2,4) | 8 |
| [4,6) | 217 |
| [6,8) | 1359 |
| [8,10) | 3363 |
| [10,12) | 3437 |
| [12,14) | 1376 |
| [14,16) | 229 |
| [16,18) | 11 |



# Drawing a distribution

## Continuous r.v.

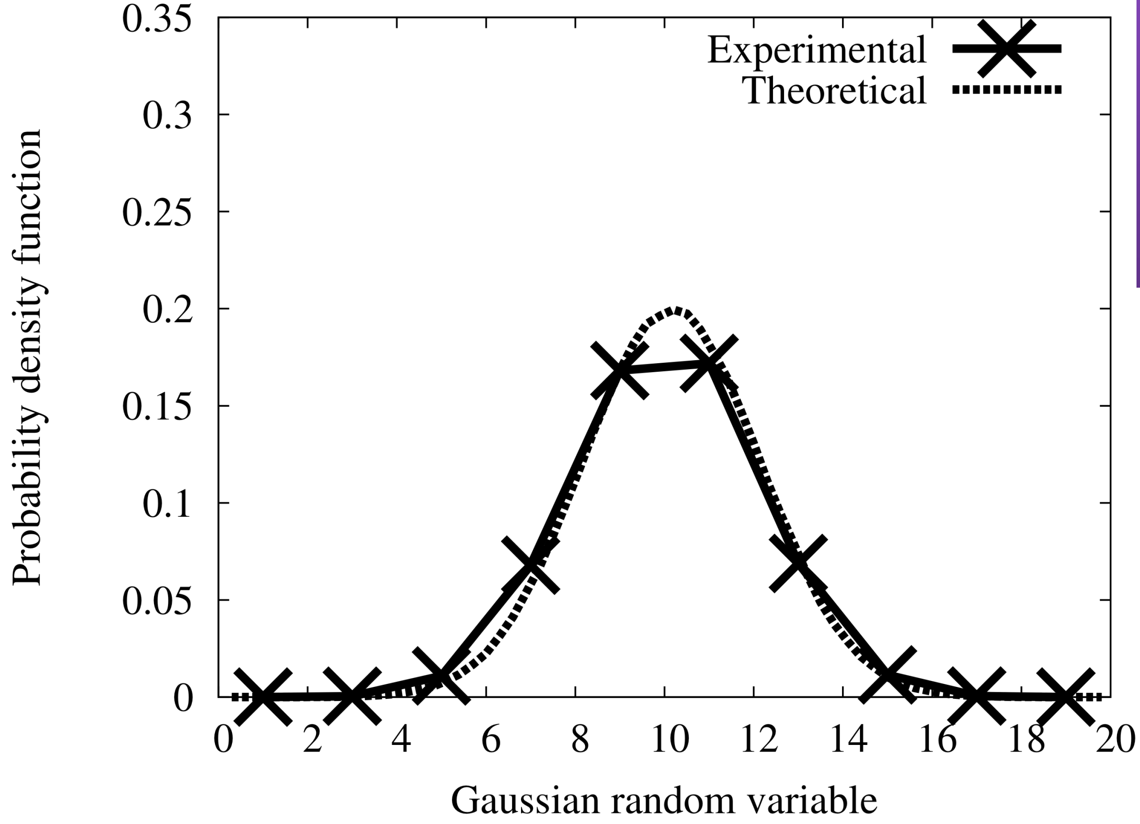
- Test: let's plot also the theoretical density function
- What has happened?!!
- In continuous rr.vv. the probability is in the AREA
- Divide also by cube width



# Drawing a distribution

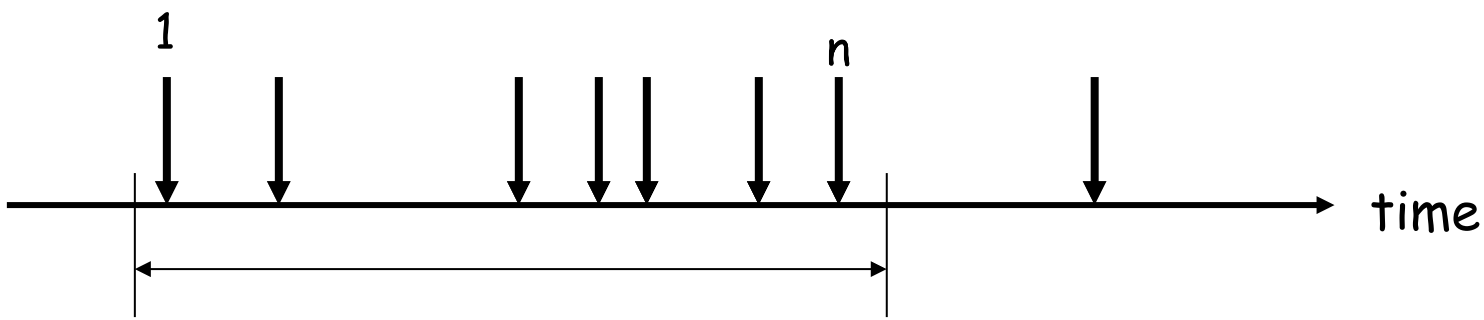
**Continuous r.v.**

- Test: let's plot also the theoretical density function
- What has happened?!!
- In continuous rr.vv. the probability is in the AREA
- Divide also by cube width



# Poisson process

- Imagine for example the random event of e-mail arrivals to a mail server
- Requirements:
  - The probability of 2 or more arrivals in a small enough time interval is 0 (only 0-1 arrivals in a small enough interval)
  - The number of arrivals in non-overlapping intervals are independent for all intervals
  - The probability of exactly 1 arrival in a small enough time interval $\Delta t$ is directly proportional to the interval width ($p=\lambda\Delta t$)
- The result is called a Poisson process

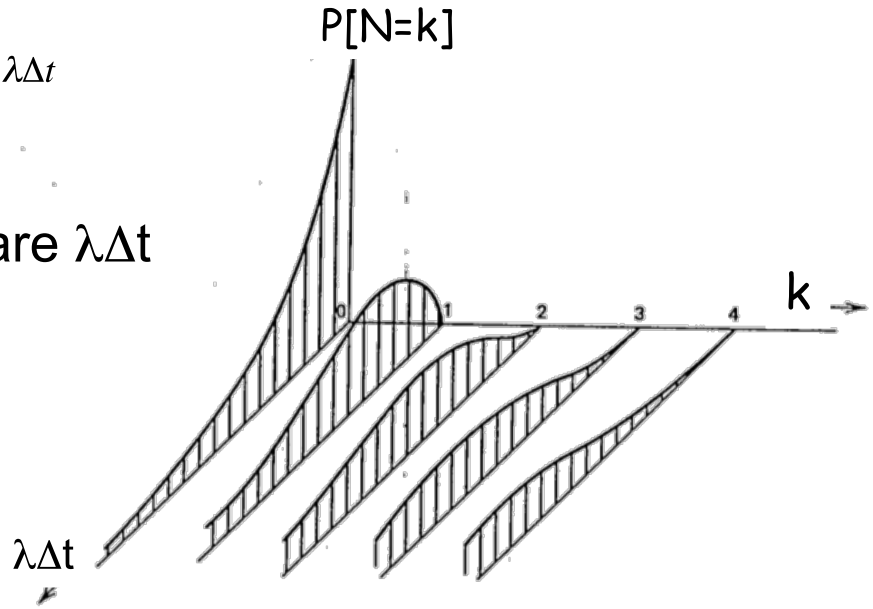


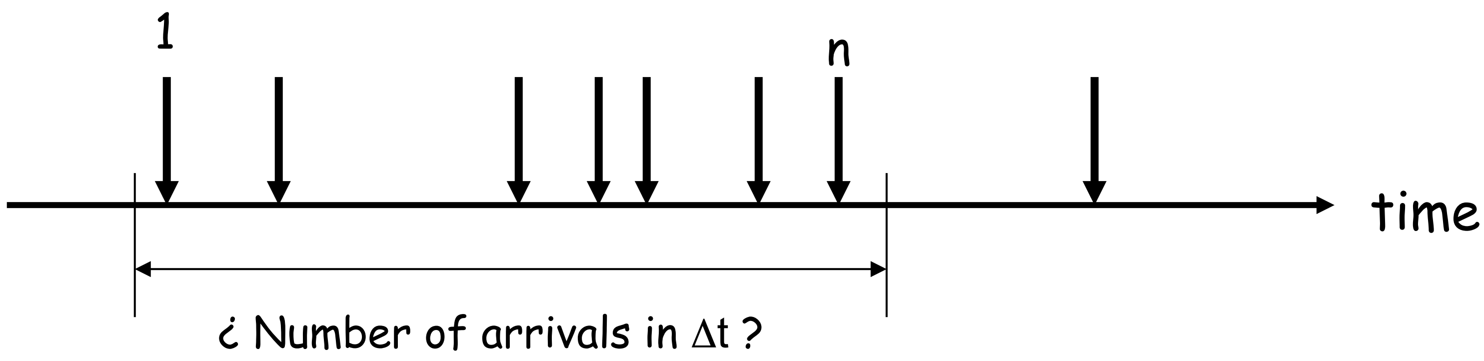
# Poisson process

- The number of arrivals in a time interval is a r.v. with a Poisson distribution:

$$P_{\lambda\Delta t}[N=k]=\frac{(\lambda\Delta t)^k}{k!}e^{-\lambda\Delta t}$$

- Expectation and variance are $\lambda\Delta t$



# Inter-arrival times

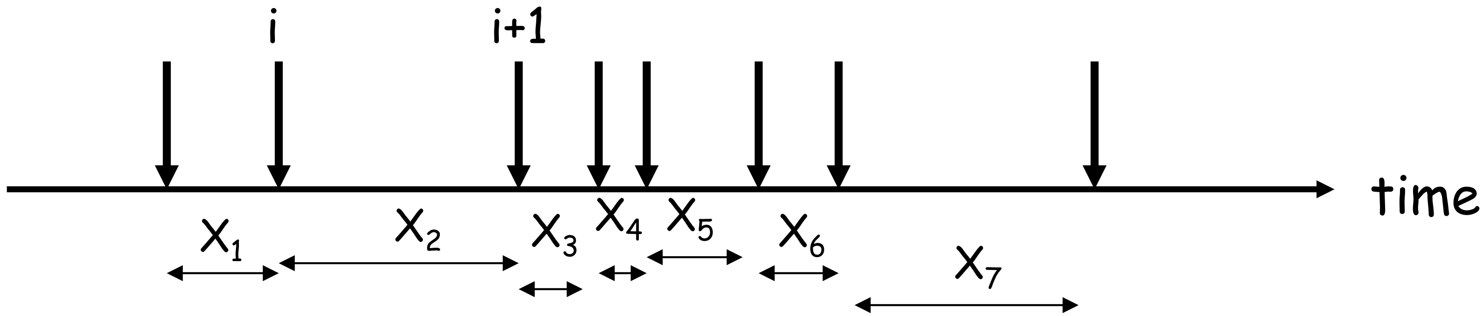
- Let $X_i$ be the time between two consecutive arrivals i and i+1
- $X_i$ are exponential i.i.d. rr.vv. iff the process is a Poisson process

$$p_{X_i}(t) = \lambda e^{-\lambda t} \quad (t>0)$$

$$P[X_i < t] = 1 - e^{-\lambda t}$$

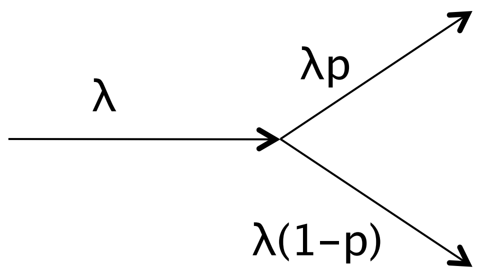
- Expectation: $E[X_i] = \int_0^{\infty} t\lambda e^{-\lambda t} = 1/\lambda$

- $1/\lambda$ is the average time between 2 consecutive arrivals ➔ there is an average of λ arrivals per time unit
- Memoryless: The probability of a future arrival in a time interval of length *s* is independent of the time of the last arrival.



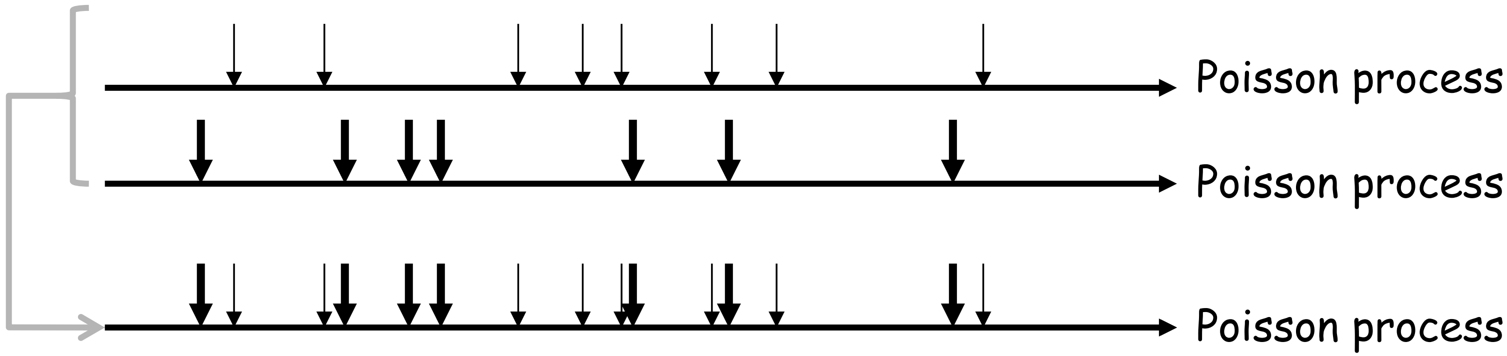
# Random splitting

- A Poisson process with rate λ
- It is split using probability p (independent)
- Resulting processes are Poisson processes with rates λp and λ(1-p)



# Limit for superposition of processes

- The superposition of two Poisson processes is a Poisson process with the aggregated rate




- For some common types of processes the superposition of a large number of i.i.d. stationary processes has a Poisson process limit