CPE 213 Data Models

(a.k.a. Data Modeling and Visualization)

Lecture 7: Modeling statistical distribution

Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi

Objectives

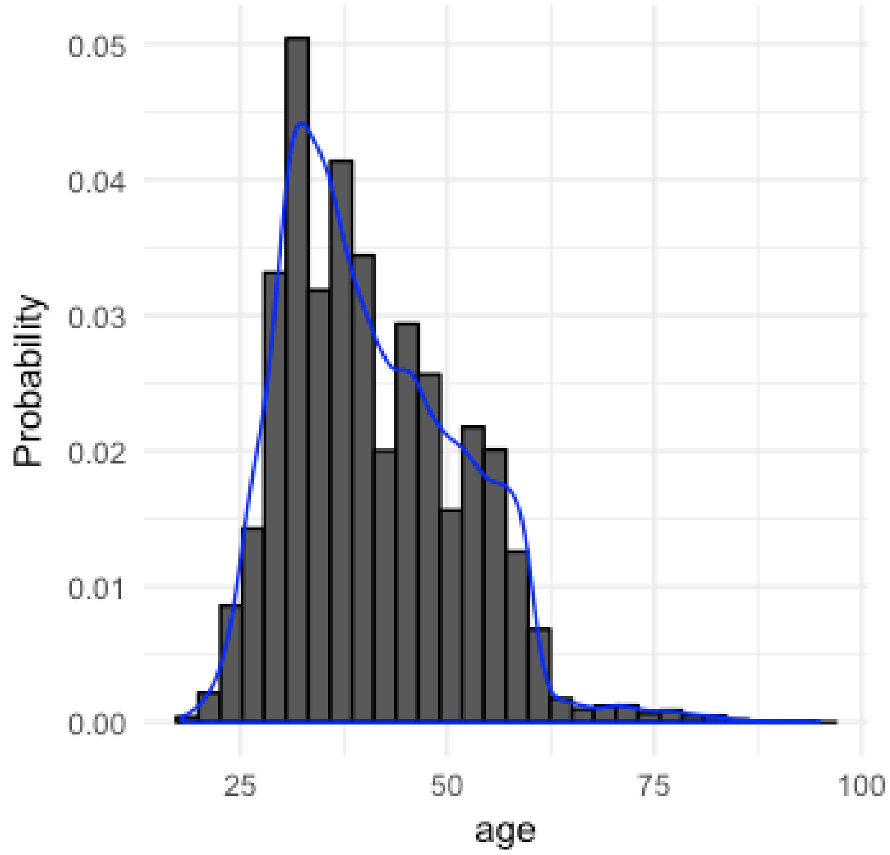
- The world the model-builder sees is probabilistic rather than deterministic.
  - Some statistical model might well describe the variations.

- An appropriate model can be developed by sampling the phenomenon of interest:
  - Select a known distribution through educated guesses
  - Make estimate of the parameter(s)
  - Test for goodness of fit

Probability distribution

A probability distribution is a function that describes the likelihood of obtaining the possible values that a random variable can assume.

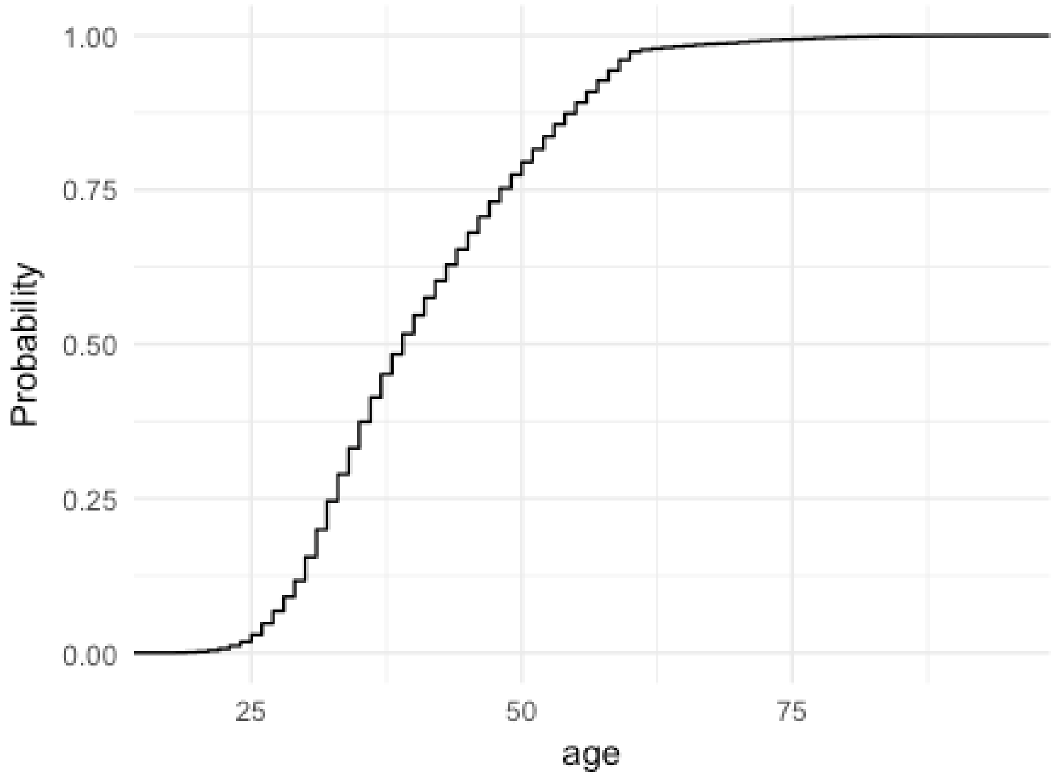
Distribution from data is called empirical distribution



Probability distribution of customer age

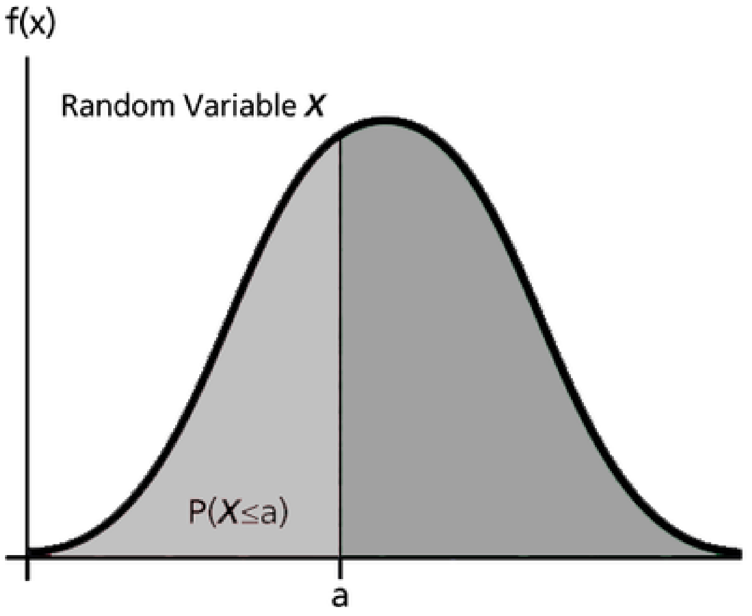
Cumulative probability distribution

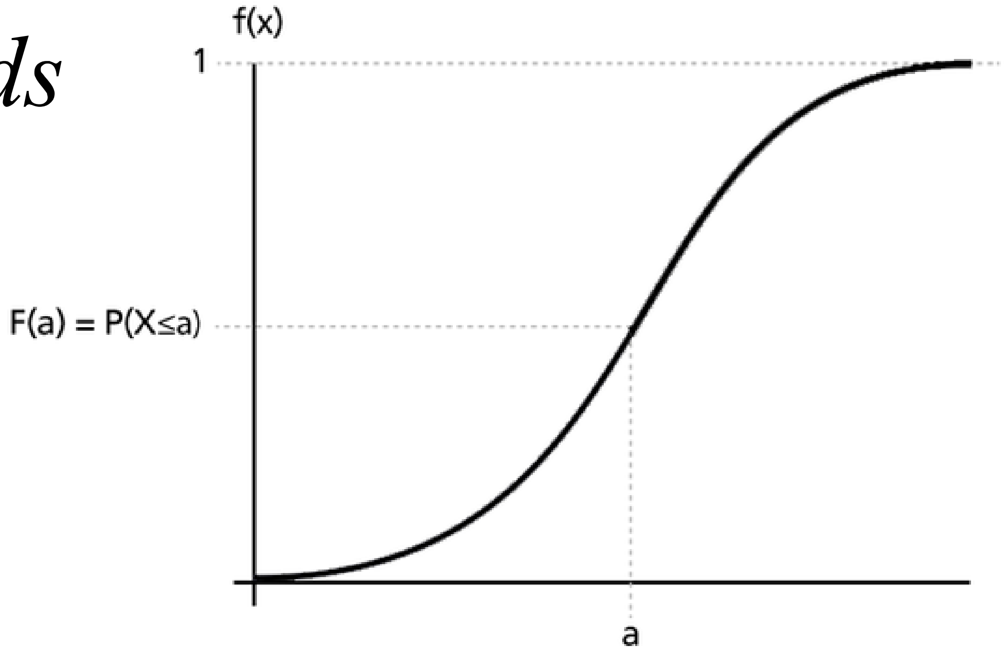
- The cumulative distribution function (CDF) $F(x)$ describes the probability that a random variable X with a given probability distribution will be found at a value.



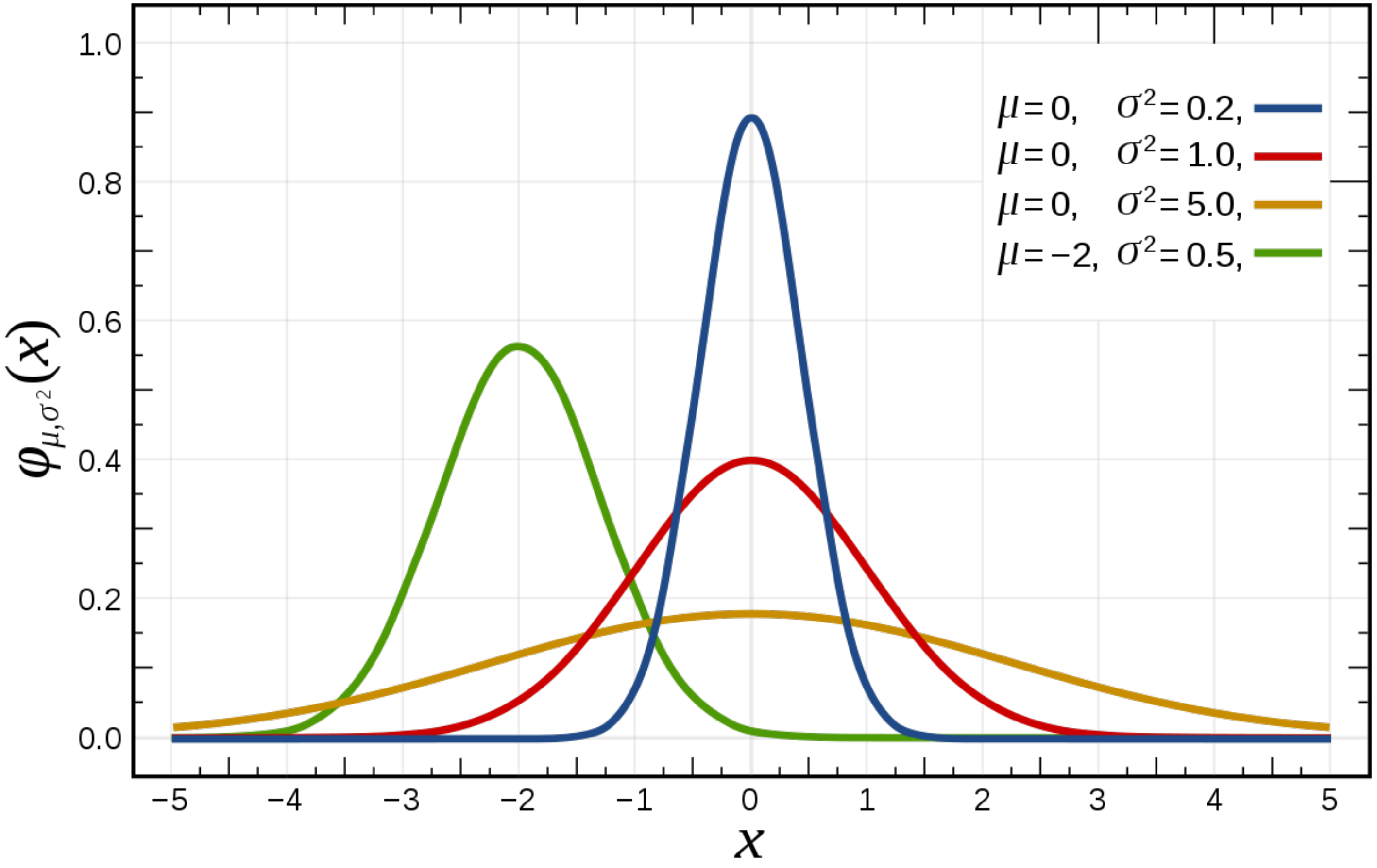
PDF and CDF relationship

$$F(x) = P(X \le x) = \int_0^x f(s)\,ds$$



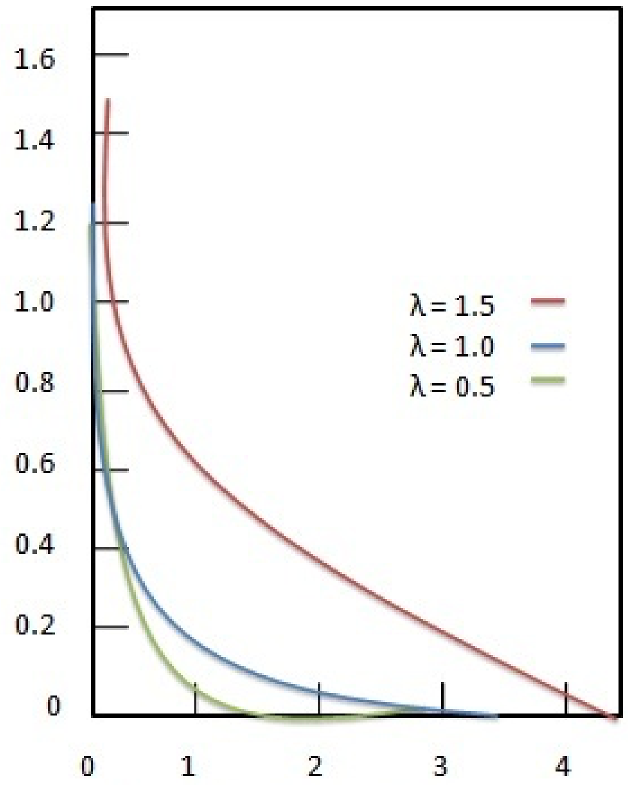


Some common distributions (PDF)
Normal

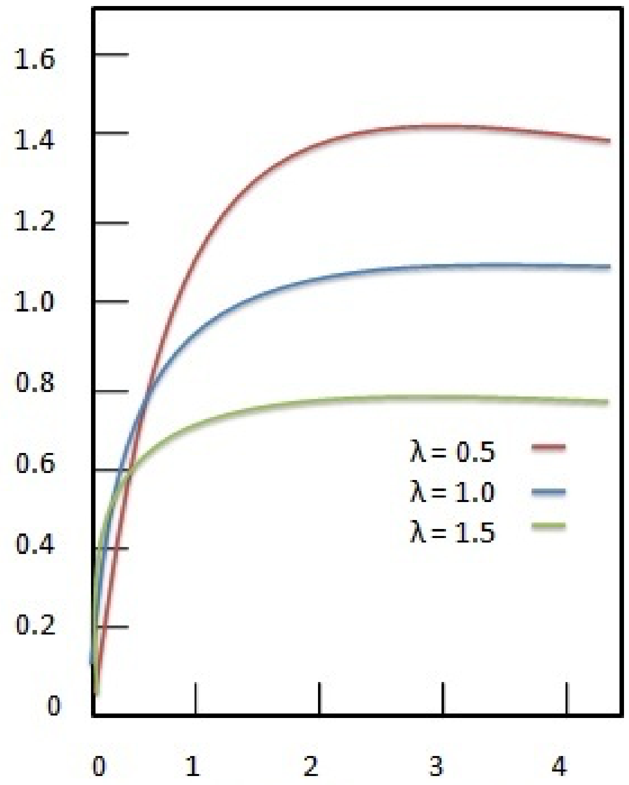


Some common distributions (PDF) Exponential

$$f(x;\lambda)=\begin{cases}\lambda e^{-\lambda x}, & x\geq 0\\ 0, & x<0\end{cases}$$



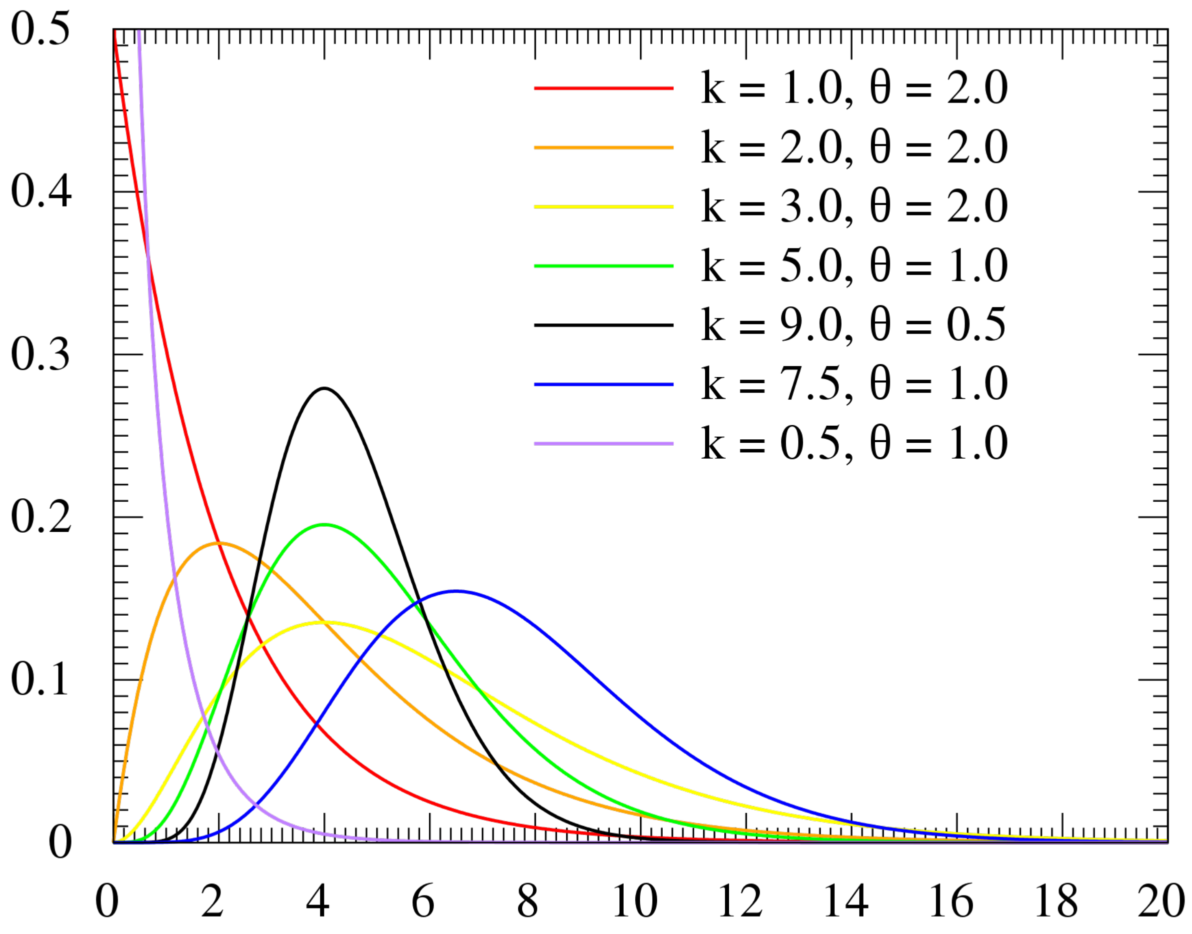
Probability Density Function



Cumulative Distribution Function

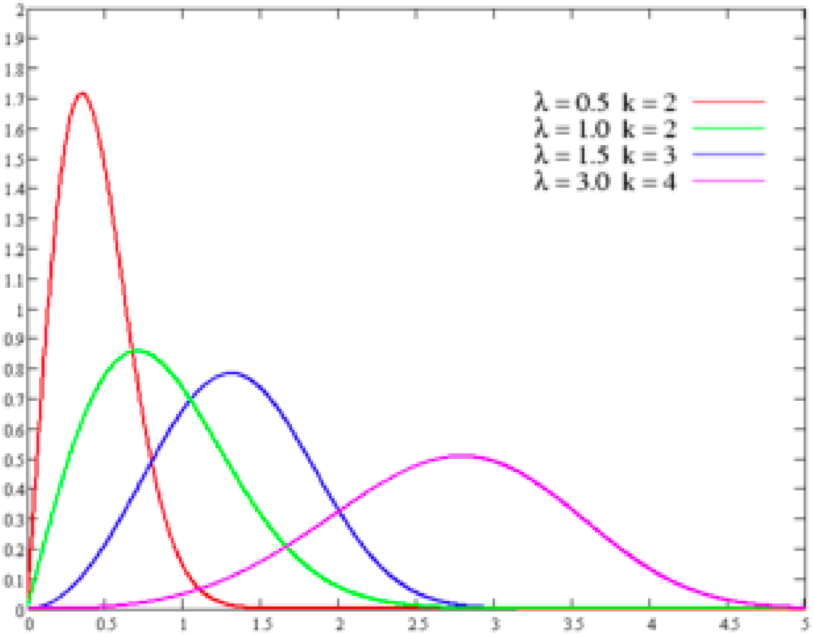
Some common distributions (PDF)
Gamma

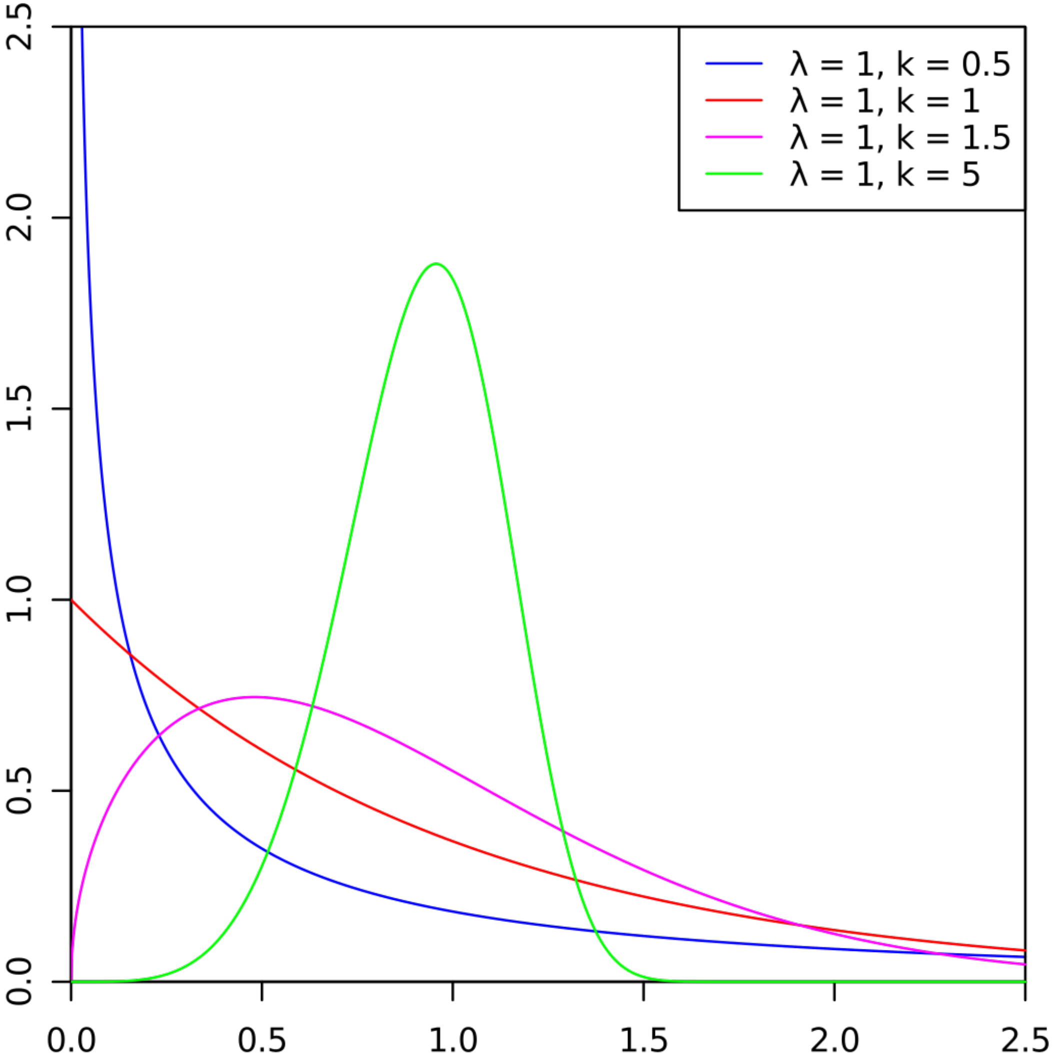
k: shape parameter
θ: scale parameter



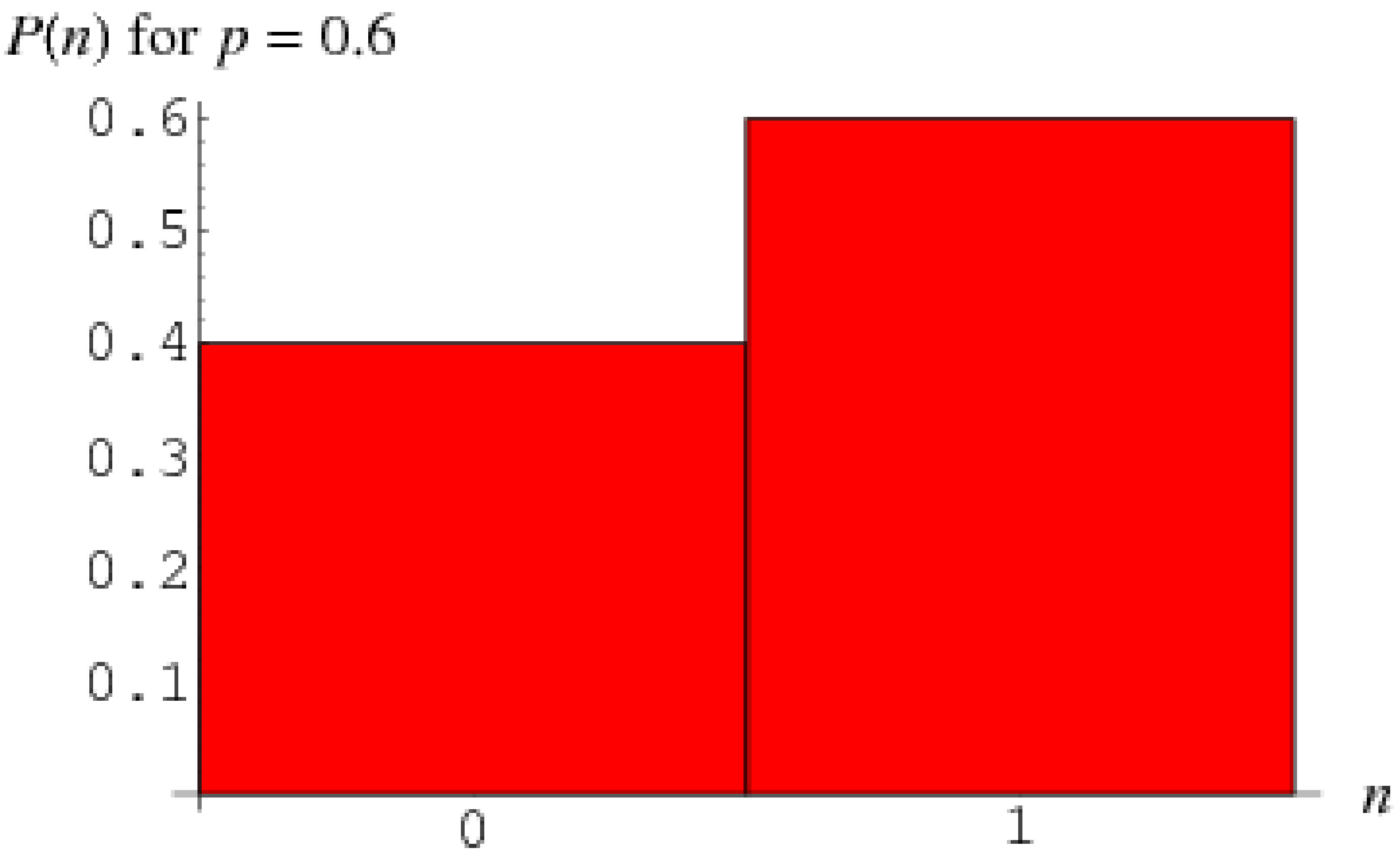
Some common distributions (PDF)
Weibull

k: shape parameter
λ: scale parameter

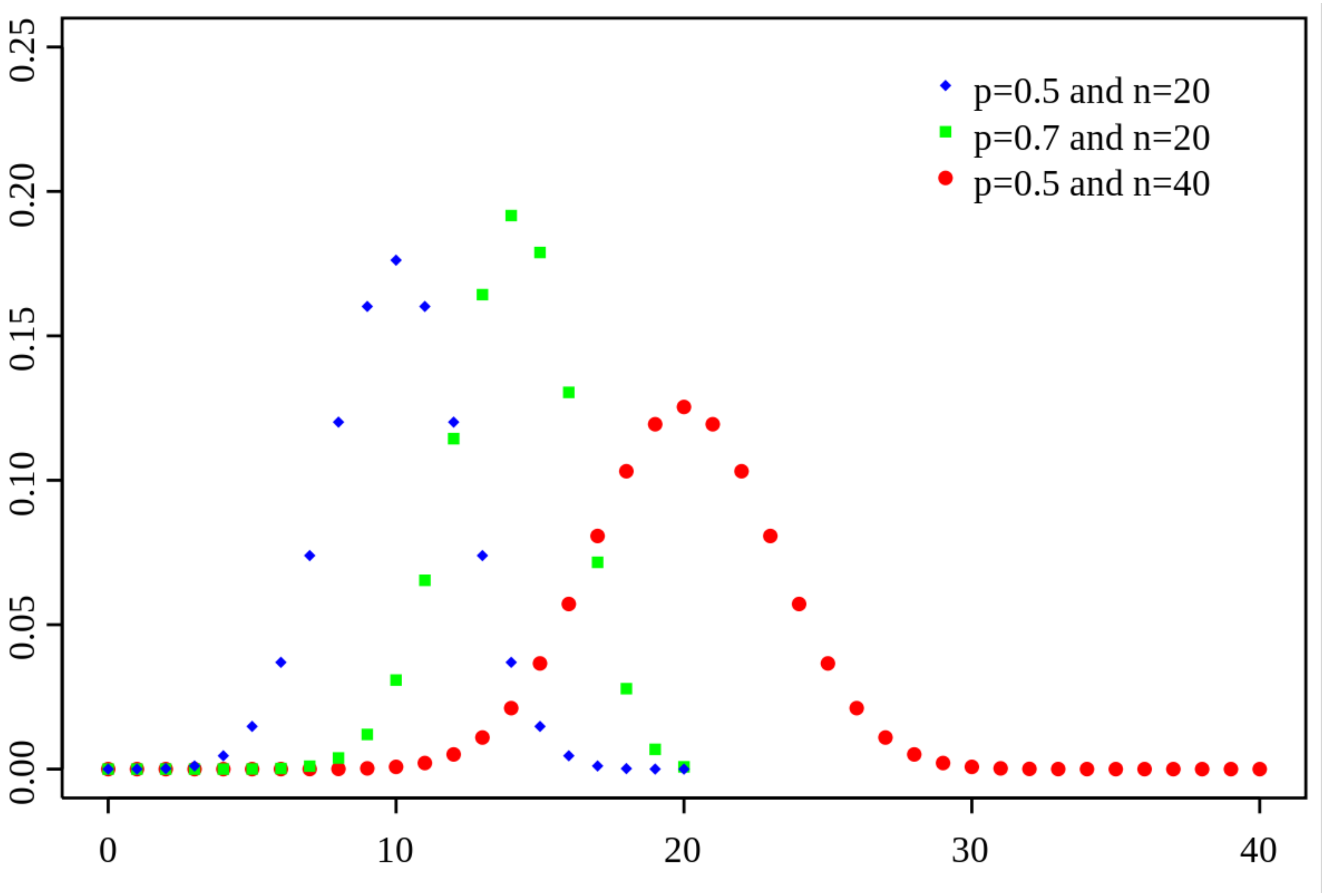




Some common distributions (PDF)
Bernoulli

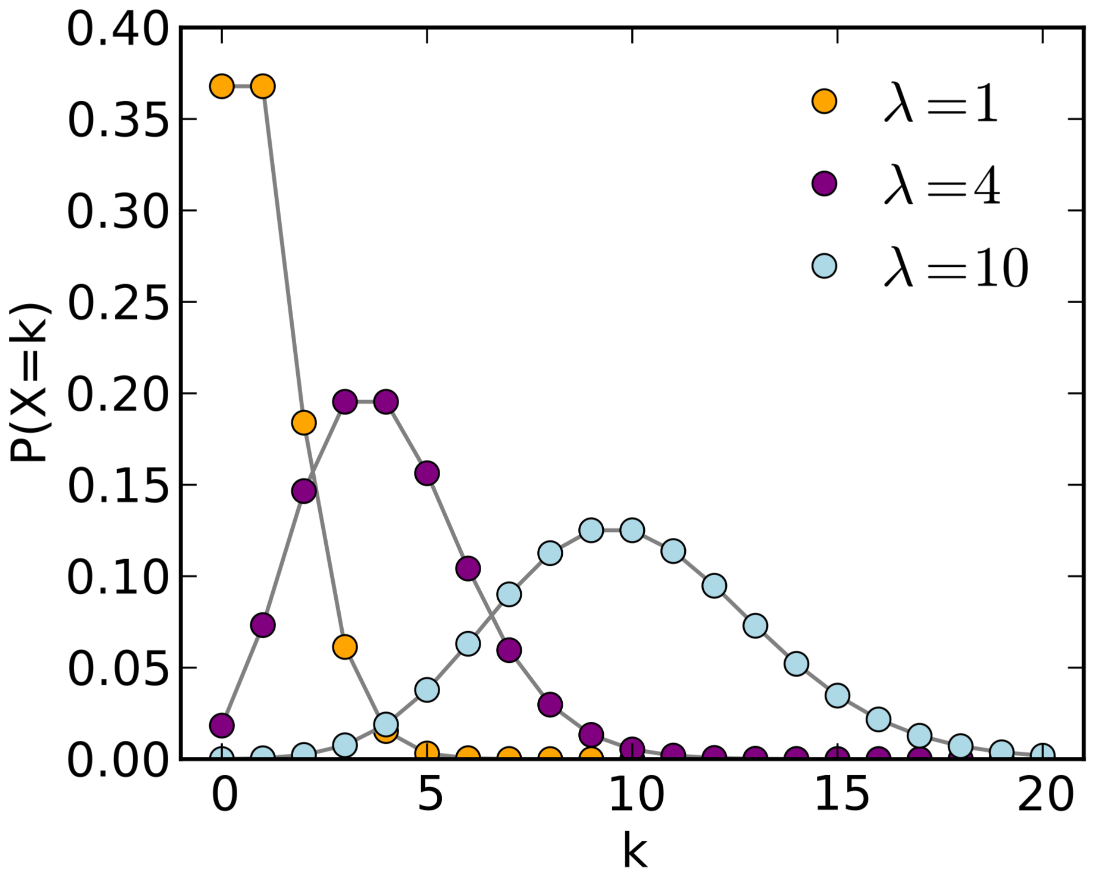


Some common distributions (PDF)
Binomial



Some common distributions (PDF)
Poisson

Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time or space



Sampling theoretical distribution

Uniform: runif

Discrete Uniform: rdunif

Normal: rnorm

Exponential: rexp

Gamma: rgamma

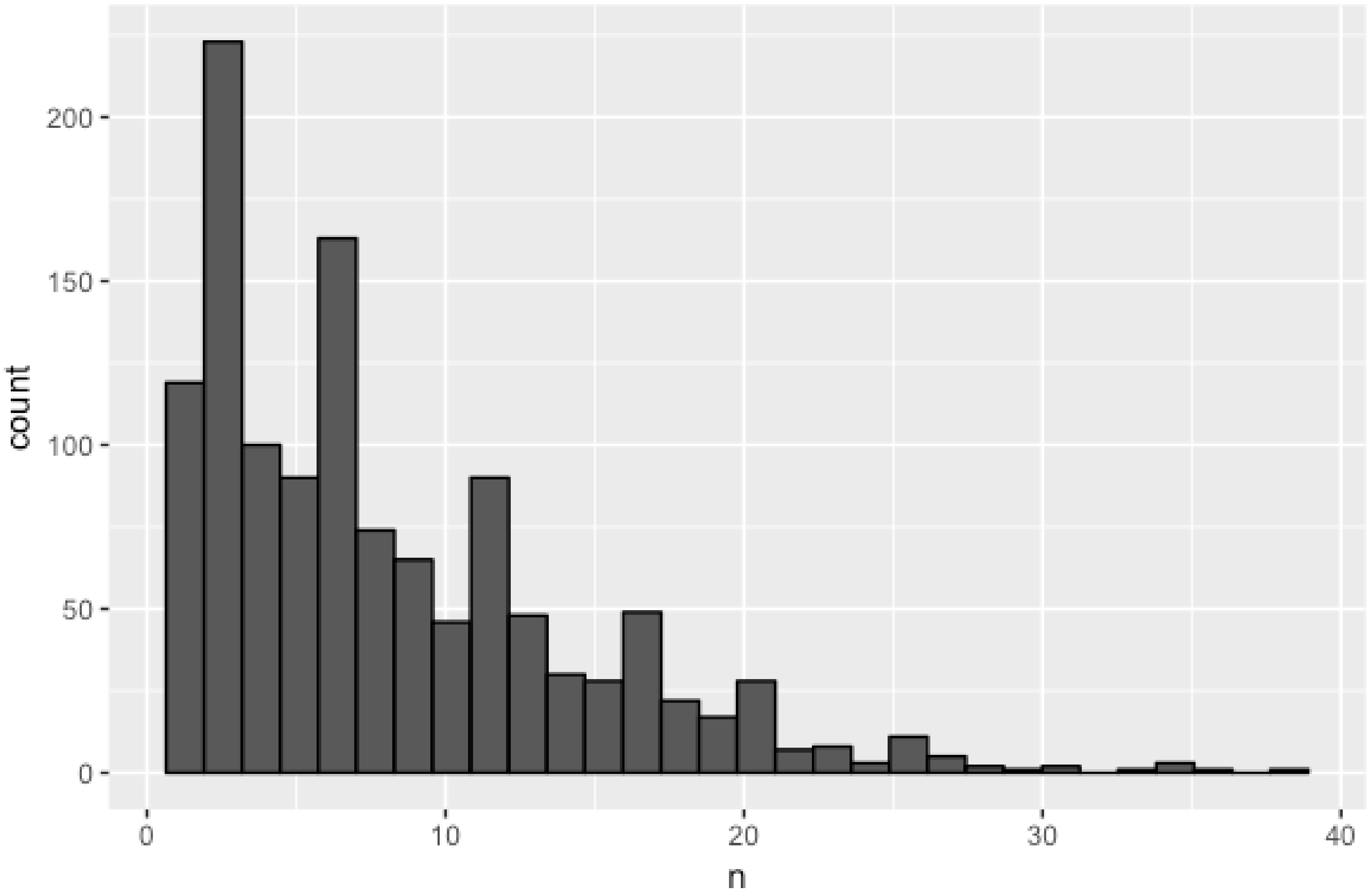
Bernoulli (Binomial): rbernoulli

Poisson: rpois

Empirical distribution

- Empirical distribution is the distribution from data

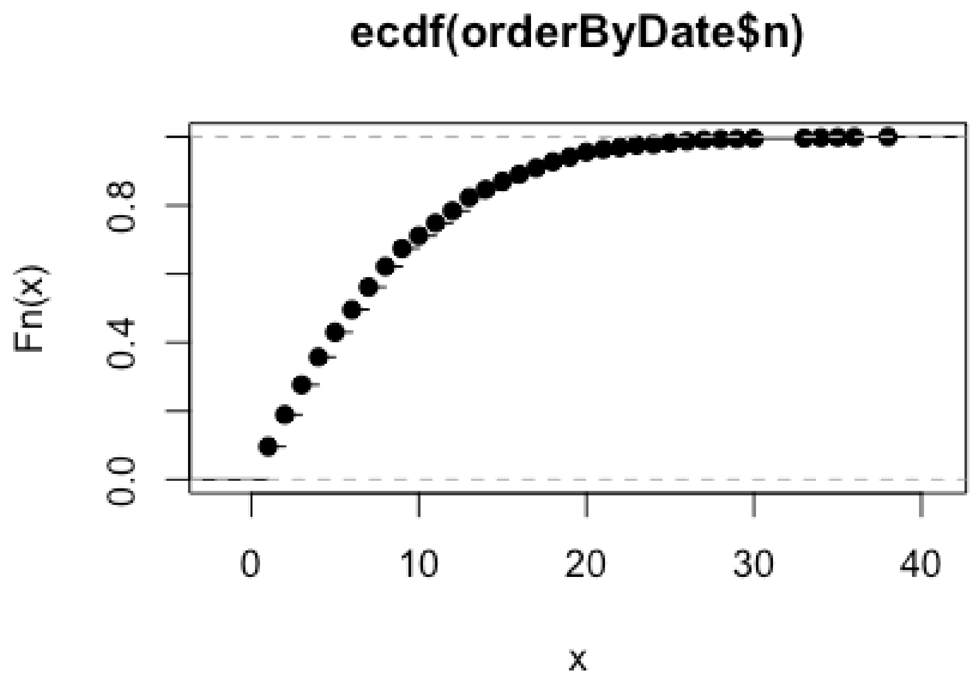
```
library(tidyverse)
library(readxl)
superstore <- read_xlsx('M2_Superstore.xlsx',
                        sheet = 1)
superstore %>%
  group_by(`Order Date`) %>%
  summarise(n = n()) -> orderByDate
ggplot(orderByDate) +
  geom_histogram(aes(x = n), color = 'black')
```



200
150
100
50
0
count
0
10
20
30
40
n

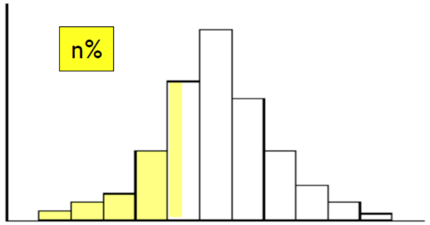
Getting empirical CDF

```
orderCDF <- ecdf(orderByDate$n)
plot(orderCDF)
```



Sampling empirical distribution Inverse ecdf is quantile

quantile(orderByDate$n, runif(10))




```
> quantile(orderByDate$n, runif(10))
25.81951% 23.45884% 4.673193% 31.12619% 11.75712%
        3         3         1         4         2
69.01843% 2.589103% 24.82679% 50.06675% 22.16411%
       10         1         3         7         3
```

Use case

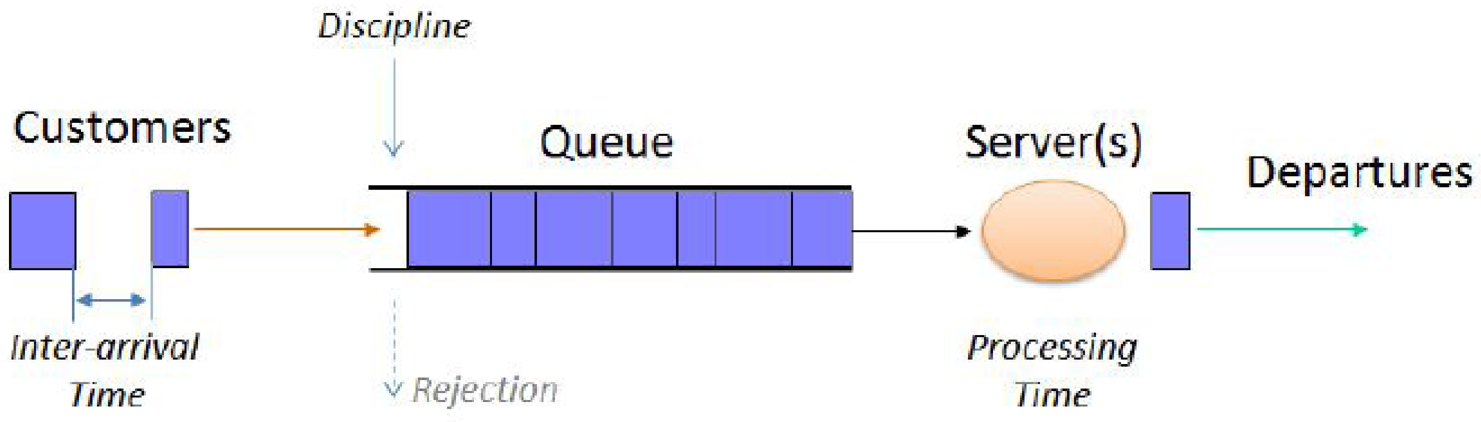
- In this section, statistical models appropriate to some application areas are presented.

The areas include:

- Queueing systems
- Inventory and supply-chain systems
- Reliability and maintainability
- Limited data

Queueing system

- In a queueing system, interarrival and service-time patterns can be probabilistic

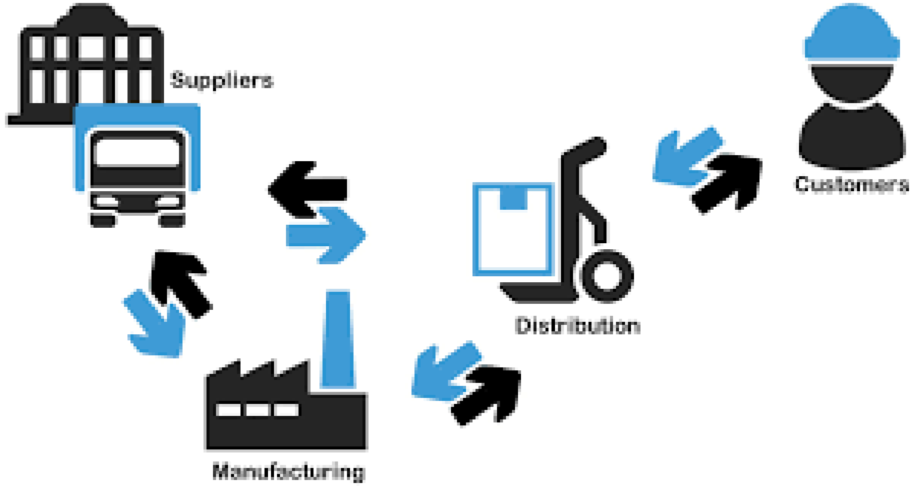


Sample statistical models for interarrival or service time distribution

- Exponential distribution: if service times are completely random
- Normal distribution: fairly constant but with some random variability (either positive or negative)
- Truncated normal distribution: similar to normal distribution but with restricted value.
- Gamma and Weibull distribution: more general than exponential (involving location of the modes of pdf's and the shapes of tails.)

Inventory and supply chain

- In realistic inventory and supply-chain systems, there are at least three random variables:
- The number of units demanded per order or per time period
- The time between demands
- The lead time



Statistical distribution models

- Sample statistical models for lead time distribution:
  - Gamma
- Sample statistical models for demand distribution:
  - Poisson: simple and extensively tabulated.
  - Negative binomial distribution: longer tail than Poisson (more large demands).
  - Geometric: special case of negative binomial given at least one demand has occurred.

Reliability and maintainability

- Time to failure (TTF)
  - Exponential: failures are random
  - Gamma: for standby redundancy where each component has an exponential TTF
  - Weibull: failure is due to the most serious of a large number of defects in a system of components
  - Normal: failures are due to wear

Summary

- The world that the simulation analyst sees is probabilistic, not deterministic.

- In this lecture:
  - Reviewed several important probability distributions.
  - Showed applications of the probability distributions in a simulation context.

Lab

- Select one numeric column of your data
- Plot probability distribution function
- Plot CDF
- Sampling 10 values from your distribution

BIG DATA EXPERIENCE

Thank you

Question?

KMUTT g-able