Statistics and Exploratory Data Analysis

Marcin Chlebus, Damian Zięba

Faculty of Economic Sciences
University of Warsaw

Lecture 1

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Organization Issues

- The goal of this class is to provide you with **theoretical knowledge** concerning basic concepts used in statistics and explanatory data analysis as well as **tools for its implementation** (using R)

- Each meeting will consists of 3-4 parts:
  - 8 minutes presentation of an article
  - Theoretical lecture
  - Practical examples in R
  - Exercises (own work)

- Your presence is **mandatory**

- In order to pass the course you need to:
  - **pass a written open book exam (80% of final grade)**
  - **in groups of 3 - present results from an article of your choice, in which methods presented at the course were used (20% of final grade)**

- Class materials will be available for you on the Google Drive.

What is statistics and why to study it?

- Statistics is a collection of methods which help us to describe, summarize, interpret, and analyse data (Heumann, 2016)
- Statistics can be seen in everyday life (e.g. unemployment rates, political party support, football scores, stock indexes, etc.)
- Statistics is used not only to inform us but also to influence us:
  - *„There are three kinds of lies: lies, damned lies and statistics"*
  - *„Statistics don't lie, but liars use statistics"*
- Efficient and inteligent use of statistics will help you to better understand and interpret everyday processes (not only in you prefessional career!)

What is EDA and why it is important?

- EDA was promoted by the statistician John Tukey in his 1977 book, “Exploratory Data Analysis”.
- **EDA is a necessary step before any more complex analysis of data** (e.g. econometric model).
- It helps to formulate hypotheses, choose the appropriate model, verify its assumptions.
- EDA involves a mix of both numerical and visual methods of analysis.

Basic concepts

- **Observations (ω)** - units in which we measure data (e.g. persons, cars, days, households, countries, etc.)

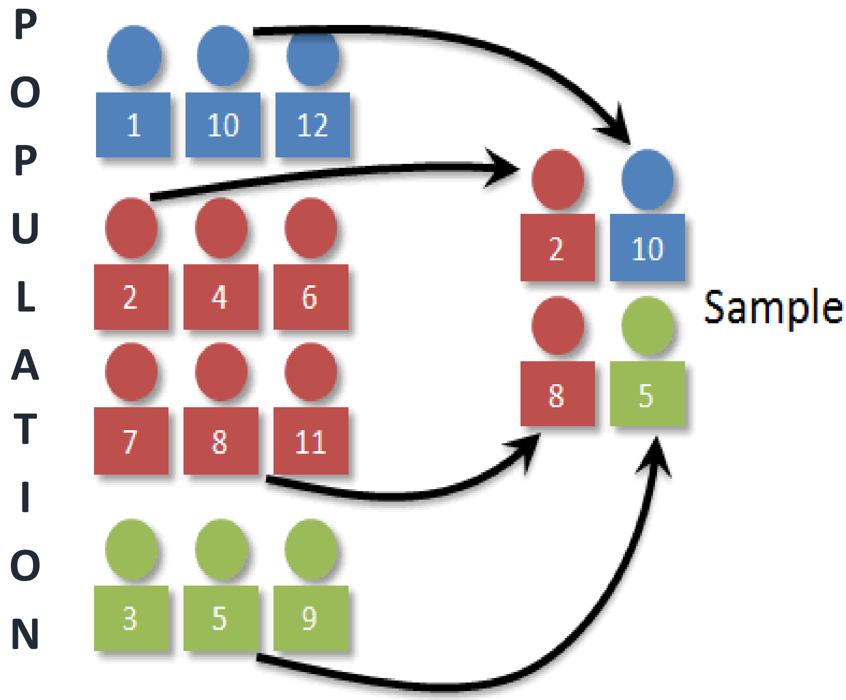
- **Population (Ω)** – the collection of all units/observations

- **Sample** $(\boldsymbol{\omega_1}, \boldsymbol{\omega_2}, \dots, \boldsymbol{\omega_n})$ - a selection of observations. A sample is always a subset of the population: $\{\omega_1, \omega_2, \dots, \omega_n\} \subseteq \Omega$

- **Variables (X)** – features of observations (e.g. grades, sex, color, location, etc.). X takes a value of x for each observation $\omega \in \Omega$, and the number of possible values is contained in the set S.

$$\begin{aligned} X : \Omega &\to S \\ \omega &\mapsto x \end{aligned}$$

Basic concepts



A number that summarizes variation in the whole population

PARAMETER

A number that summarizes variation in the subset of a population (sample)

STATISTICS

http://faculty.elgin.edu/dkernler/statistics/ch01/1-4.html

Basic concepts

Observations: students

Variables: courses they participated in, grades

STUID	STUNAME	COURSENUM	GRADE
S1010	Burns,Edward	MTH103C	A
S1010	Burns,Edward	ART103A	B
S1002	Chin,Ann	ART103A	A
S1002	Chin,Ann	MTH103C	B
S1002	Chin,Ann	CSC201A	F
S1020	Rivera,Jane	MTH101B	A
S1020	Rivera,Jane	CSC201A	B
S1001	Smith,Tom	HST205A	C
S1001	Smith,Tom	ART103A	A

Sample: 4 students

Population: students

Types of variables

- **Qualitative** - variables which take values that cannot be ordered in a logical or natural way (e.g. sex, color, name of a political party, taste, etc.)

  → it is common to assign numbers to qualitative variables for practical purposes in data analyses

  → Such variables are usually stored in R as **factors**

- **Quantitative** - variables which take values which can be ordered in a logical and natural way (measurable quantities, e.g. price, size, lenght, height, weight, etc.)

  → Such variables are usually stored in R as **integers** or **numeric**

Types of variables

- **Discrete** - variables which can only take a finite number of values.

  All qualitative variables are discrete; quantitative variables can be discrete.

- **Continuos** - variables which can take an infinite number of values.

  Informally, continuous variables are variables which are “measured rather than counted”.

Scales

- **Nominal scale** - the values of a nominal variable cannot be ordered.
  Example: gender (male–female)

- **Ordinal scale** - the values of an ordinal variable can be ordered but the differences between these values cannot be interpreted in a meaningful way.
  Example: education level, satisfaction level.

- **Continuous scale** - the values of a continuous variable can be ordered and the differences between these values can be interpreted in a meaningful way.
  Example: the height/length/weight.
  - Interval (with arbitrary 0 - temperature) and ratio (with 0 means nothing - distance) scales

Variables vs. Random variables

- Random variable is a mathematical concept, which helps us to view the collected data as an outcome of a **random experiment** (e.g. tossing a coin, randomly asking people about their grades, etc.) and that helps us to draw **conclusions formulated based on sample about the population of our interest.**

  Random variable is thus a variable whose value **is determined by a chance event.**

- Formally (recall from the Probability Theory):
  Let Ω represent the sample space of a random experiment, and let R be the set of real numbers.
  A random variable is a function X which assigns to each element $\omega \in \Omega$ one and only one number $X(\omega) = x,\ x \in R$, i.e.
  $X : \Omega \rightarrow R$.

- Random variables may be discrete (e.g. the numer of heads/tails in tossing a coin) or continuous (age of randomly selected individuals).

The distribution of a random variable

- To infer about the distribution of a random variable we may consider **probability distribution**, which is a set of probabilities defined for all the possible outcomes of a random variable.

Discrete variable	**Continous variable**
Probability functions are denoted as *p(x)* and known as **probability mass functions** (pmf) For a function $p(x_k) = P(X = x_k)$ to be a pmf of X, it needs to satisfy the following conditions: 1) $0 \leq P(X = x_i) \leq 1$ 2) $\sum_i^n P(X = x_i) = 1$	Probability functions are denoted as *f(x)* and are known as **probability density functions** (pdf) For a function *f (x)* to be a pdf of X, it needs to satisfy the following conditions: 1) $f(x) \geq 0 \; for \; all \; x \in R$ 2) $\int_{-\infty}^{+\infty} f(x)dx = 1$

Cummulative distribution function (CDF)

- For both discrete and continous variables we can also define **cummulative distribution function (cdf)**.
- The cdf shows the probability that a random **variable will be less than or equal to a specific value** for all possible values of that variable: $F(x) = P(X \leq x)$
- The cumulative distribution function is represented as:
  - **discrete random variable:** the sum of the probabilities of the specified outcome and all prior outcomes for each and every possible outcome
  - **continous random variable:** an integral over the pdf: $F(X) = \int_{-\infty}^{t} f(t)dt$

Empirical (i.e. observed) cdf

- The empirical cumulative distribution function is the estimator for the population's cumulative distribution function, which contains all the characteristic of the population.
- It is often interpreted as a graph of a **cumulative frequency.**

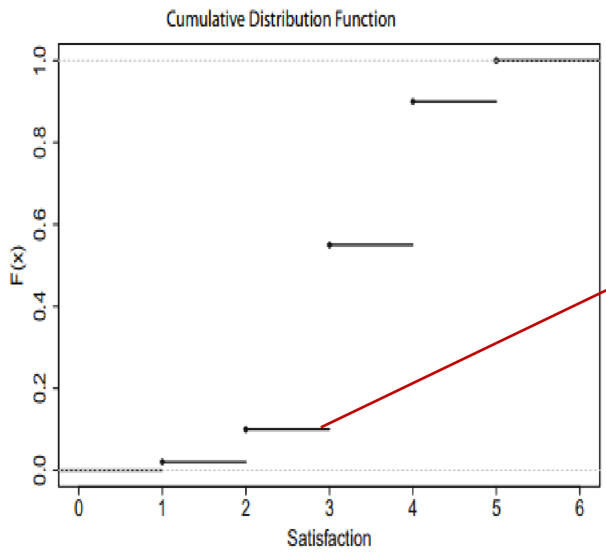
Example: discrete random variable

Suppose there are 100 randomly selected consumers who were asked about their overall level of satisfaction with the quality of pizza on a scale from 1 to 5 based on the following options: 1 = not satisfied at all, 2 = unsatisfied, 3 = satisfied, 4 = very satisfied, and 5 = perfectly satisfied. Their answers are as follows:

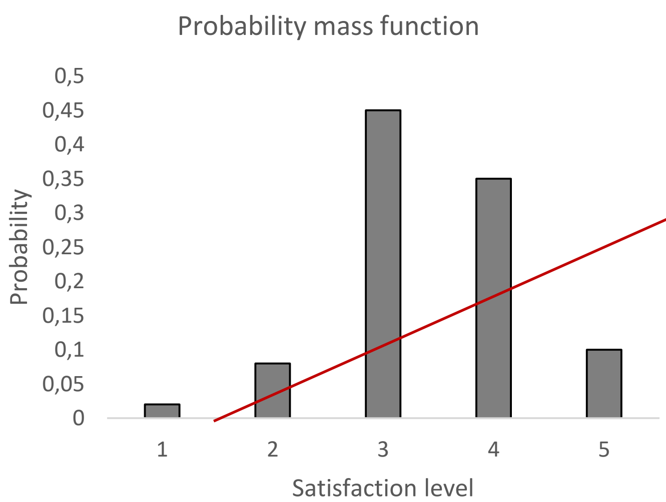
Satisfaction level (x_i)	1 = not satisfied at all	2 = unsatisfied	3 = satisfied	4 = very satisfied	5 = perfectly satisfied
N	2	8	45	35	10
P(X=x_i)	2/100	8/100	45/100	35/100	10/100
F(X)	2/100	10/100	55/100	90/100	100/100

The sum of the probabilities of the specified outcome and all prior outcomes for each and every possible outcome

Example: discrete random variable



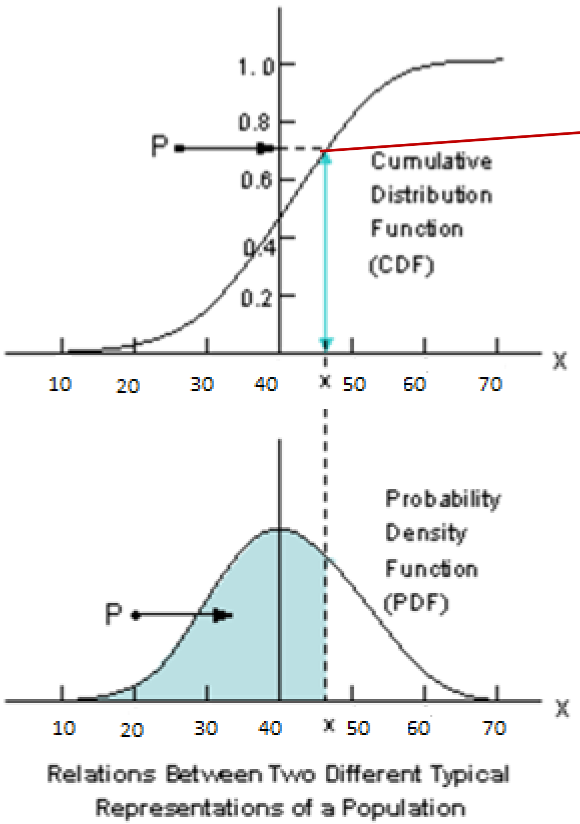
Probability that consumers are not satisfied with pizza is 10%: **10% of consumers are not satisfied with the pizza**



Probability that consumers were „not satisfied at all" is 2% and the probability that consumer were „unsatisfied" is 8%.

Example: continous random variable

Suppose there are 100 randomly selected consumer who were asked about their age.



Probability that consumers are at most 47 years old is 70%: **70% of consumers are 47 or younger.**

Source: https://home.ubalt.edu/ntsbarsh/Business-stat/opre504.htm

PDF and CDF in R

Exercise 1:

Use data on consumers' satisfaction and age from the file „pizza.csv".

- Create the pdf and cdf functions for variables satisfaction and age.

Based on pdf and cdf answer the questions:

- What is the probabiliy that the consumers are younger than 40?
- What is the share of the consumers that are younger than 40?

Cummulative distribution function (CDF)

„What are the chances that X takes some subset of values?”

- The event that X is less than or equal to ***b*** but not less than or equal to ***a*** is the event that X is greater than ***a*** and less than or equal to ***b***.
- By the difference rule for probabilities:
  $P\{a < X \leq b\} = P(\{X \leq b\} \cap \{a \leq X\}) = P\{X \leq b\} - P\{X \leq a\} \ = \boldsymbol{F}(\boldsymbol{b}) \ - \boldsymbol{F}(\boldsymbol{a})$
- We can compute the probability that a random variable takes values in an interval by subtracting the CDF evaluated at the endpoints of the intervals.

PDF and CDF in R

Exercise 1 con't:

Use data on consumers' satisfaction and age from the file „pizza.csv".

- Create the pdf and cdf functions for variables satisfaction and age.

Based on pdf and cdf answer the questions:

- What is the probabiliy that the consumers are younger than 40?
- What is the share of the consumers that are younger than 40?
- What are the chances that the consumers are between 20 and 40 years old?

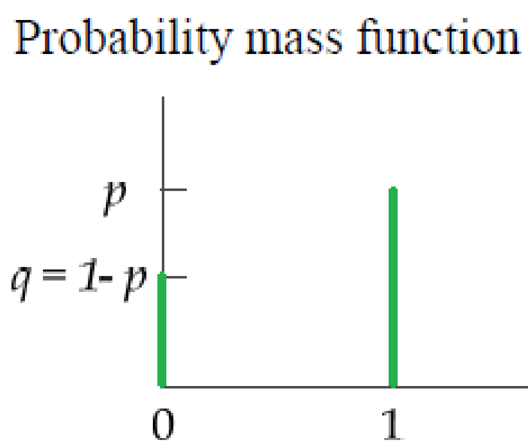
The most common distribution functions

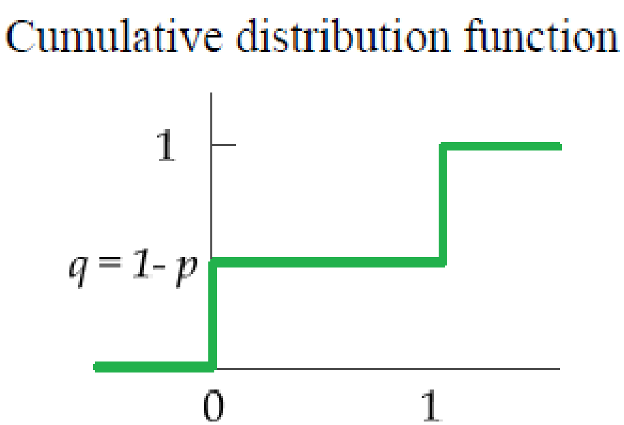
Discrete	Continous
• Bernoulli distribution	• **Normal distribution**
• Binomial distribution	• Log-normal distribution
• Geometric distribution	• Gamma distribution
• Poisson distribution	• Chi-square distribution
• And more (e.g. negative binomial distribution)	• Student's t distribution
	• And more (e.g. exponential distribution, Cauchy distribution)

Bernoulli distribution

The Bernoulli random variable takes value 1 with success probability p and value 0 with failure probability q = 1 - p.

The pmf is given by:
p for k=1
q for k=0





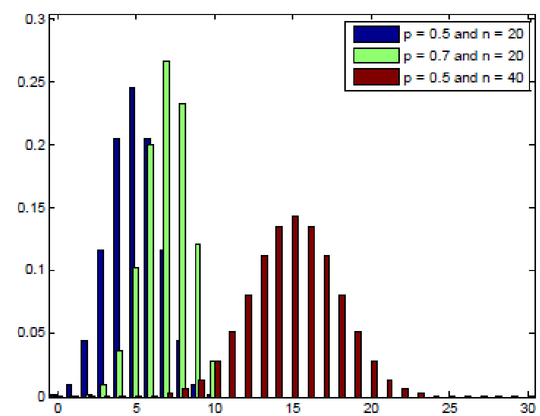
Binomial distribution

The binomial distribution describes obtaining k succeses in n experiments (e.g. obtaining k heads in n tossing of coin)

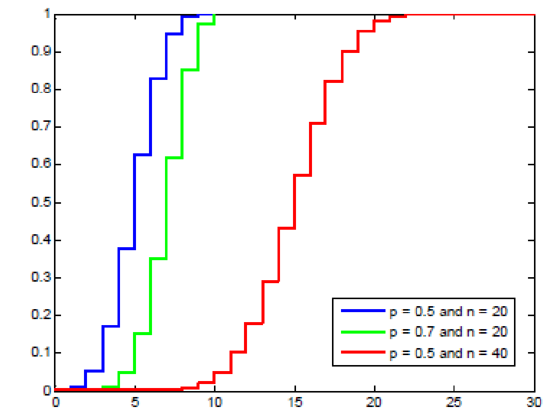
The pmf is given by:

$$P(X=x)=\binom{n}{x}p^{x}(1-p)^{n-x}$$

Probability mass function



Cumulative distribution function



Binomial distribution

Exercise 2: Binomial distribution

Calculate the probability of obtaining 0 heads in 4 coin tossing.

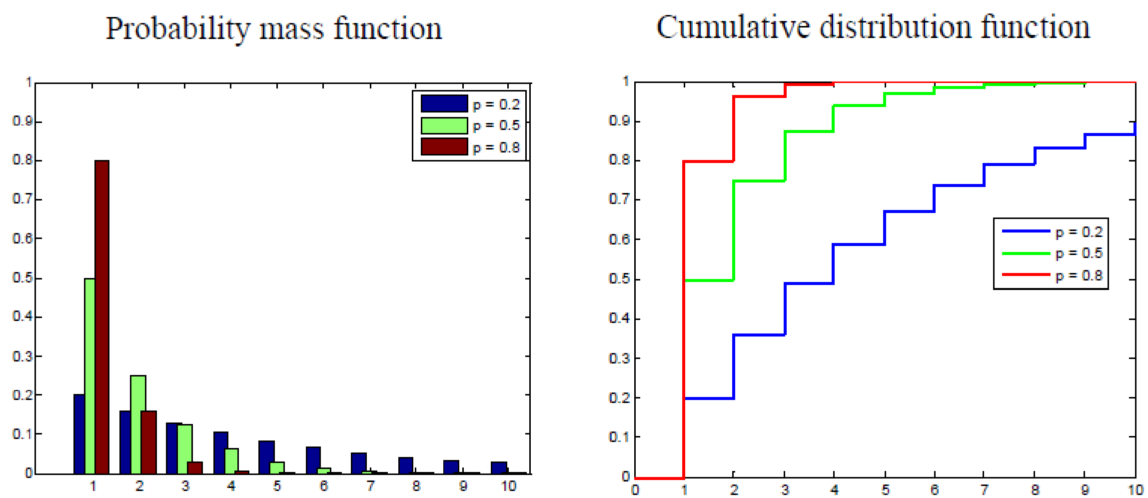
What are the chances of receiving more than 3 heads in 4 coin tossing?

Geometric distribution

The geometric distribution describes distribution of time between the successes of successive independent Bernoulli trails (e.g. the numer of coin tossing needed to obtain a head; the numer of dice rolls needed to obtain 1)

The pmf is given by:

$$(1-p)^{k-1}p$$

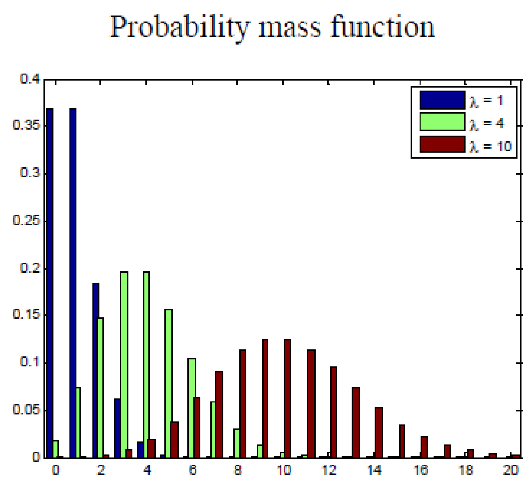


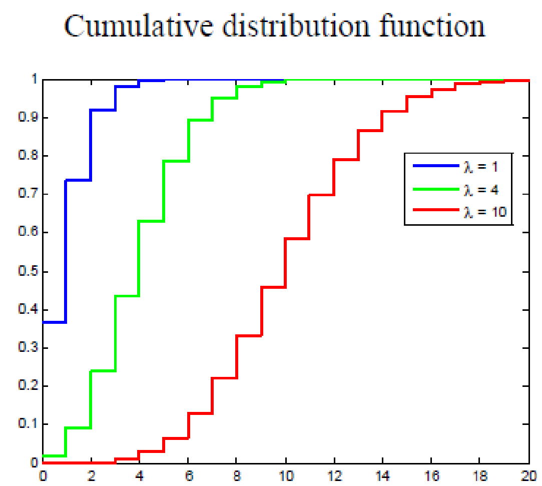
Poisson distribution

The Poisson distribution describes the probability that k events occur in a fixed time period, assuming that they appear at random with a rate λ (e.g. the numer of phone call to a call center per minute, the numer of spelling mistakes made while typing a page of a text)

The pmf is given by:

$$\frac{e^{-\lambda}\lambda^{k}}{k!}$$





Poisson distribution

Exercise 3: Poisson distribution

Consider a population of raisin buns for which there are an average of 3 raisins per bun, i.e. $\lambda = 3$.

The number of raisins in a particular bun is uncertain;
the possible numbers of raisins are 0, 1, 2, . . .

Calculate the probability of finding exactly 2 raisins in a bun.

What are the chances of finding more than 3 raisins in a bun?

Normal and standard normal distribution

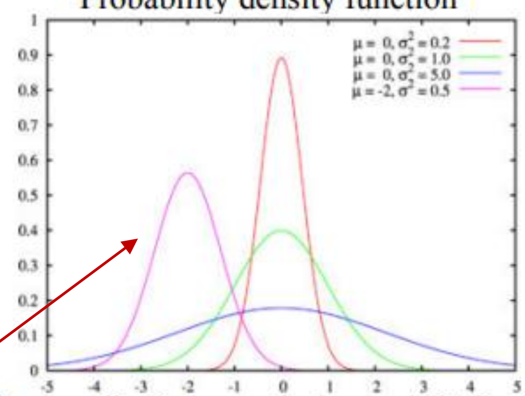
Normal distribution is also known as Gaussian distribution and is denoted by $N(\mu,\sigma^2)$.

When μ=0 and σ^2=1 it is known as st.normal distribution.

The pdf is given by:

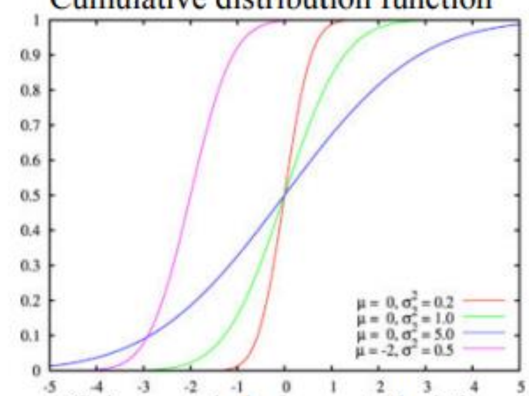
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Probability density function



The green line is the standard normal distribution

Cumulative distribution function



Colors match the image to the left

The „bell curve”

Normal and standard normal distribution

Exercise 4: Normal distribution

Consider a farmer who sells apples in wooden boxes.

The weights of the boxes vary and are assumed to be normally distributed with μ= 15 kg and σ^2 = 9/4 kg2. The farmer wants to avoid customers being unsatisfied because the boxes are too low in weight. He therefore asks the following question:

- What is the probability that a box with a weight of 10 to 15 kg is sold?

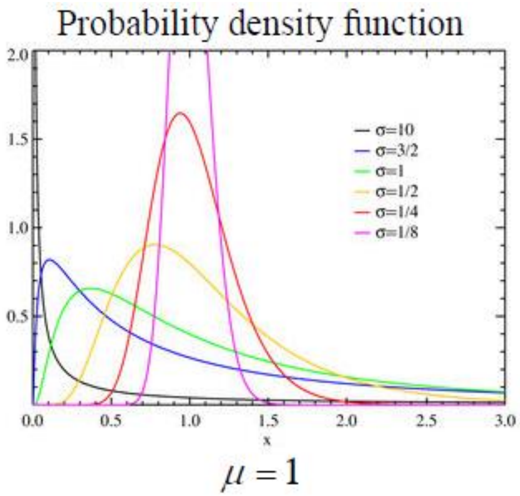
Answer this question by making relevent calculations and by using pdf and cdf graphs.

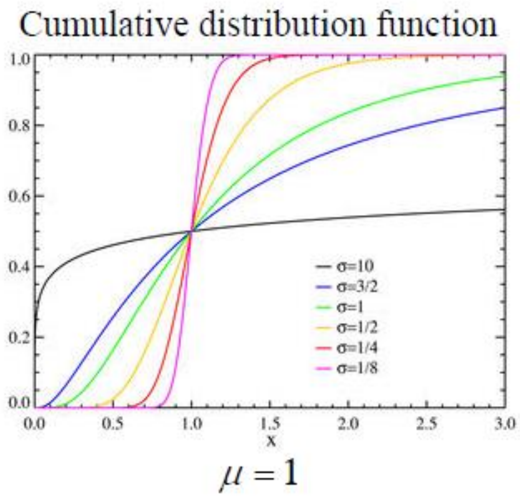
Log-Normal distribution

If Y has a normal distribution, then X=exp(Y) has a log-normal distribution (the same is true: if X has a log-normal distribution, then Y=ln(X) has a normal distribution); (e.g. household consumption, income)

The pdf is given by:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$





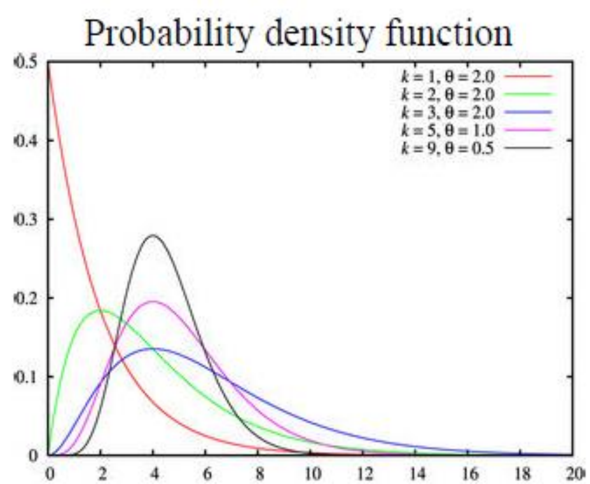
Gamma distribution

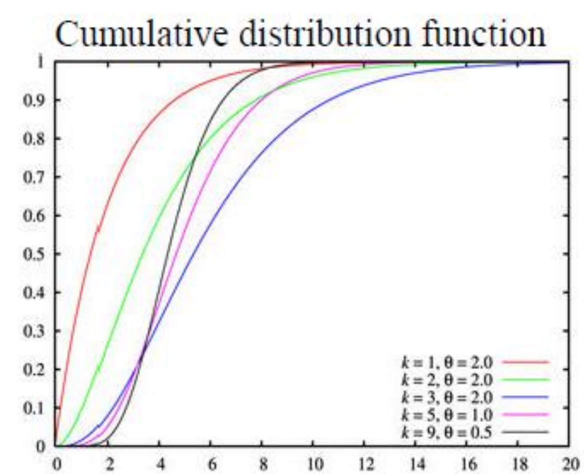
The gamma distribution is denoted by $\Gamma(k,\lambda)$ (e.g. the number of telephone calls which might be made at the same time)

The pdf is given by:

$$f(x) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\Gamma(k,\lambda)}$$

Where θ and k are parameters (θ>0 scale, k>0 shape)





Chi-square distribution

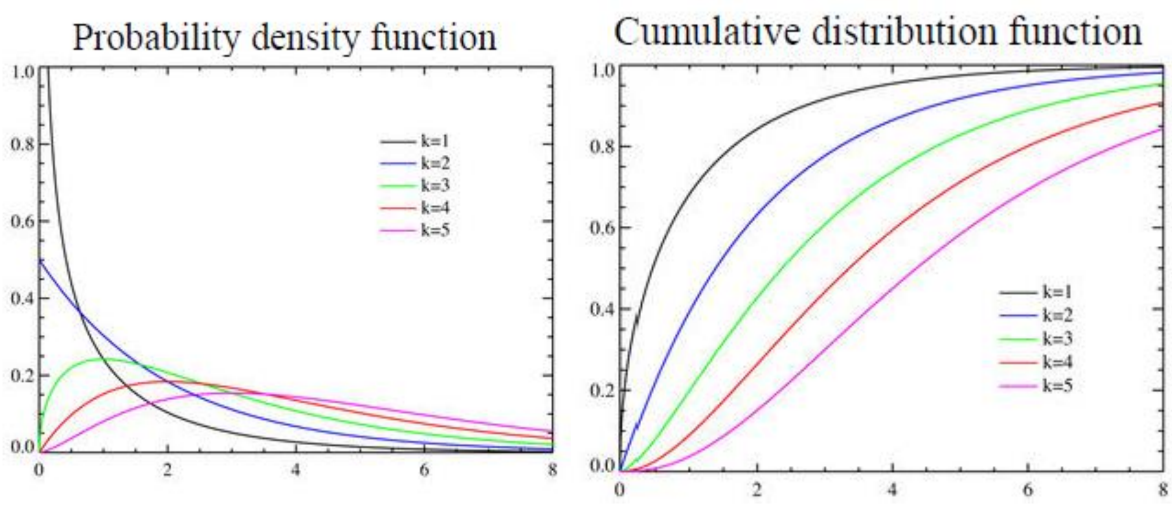
Chi-square distribution is a special case of a gamma distribution where k=v/2 and θ=2

It is one of the most widely used probability distributions in inferential statistics (goodness of fit test, independence etc.)

The pdf is given by:

$$f(x) = x^{\frac{v}{2}-1} \frac{e^{-x/2}}{2^{v/2}\Gamma(v/2,2)}$$

Where v is known as „degrees of freedom”



Student’s t distribution

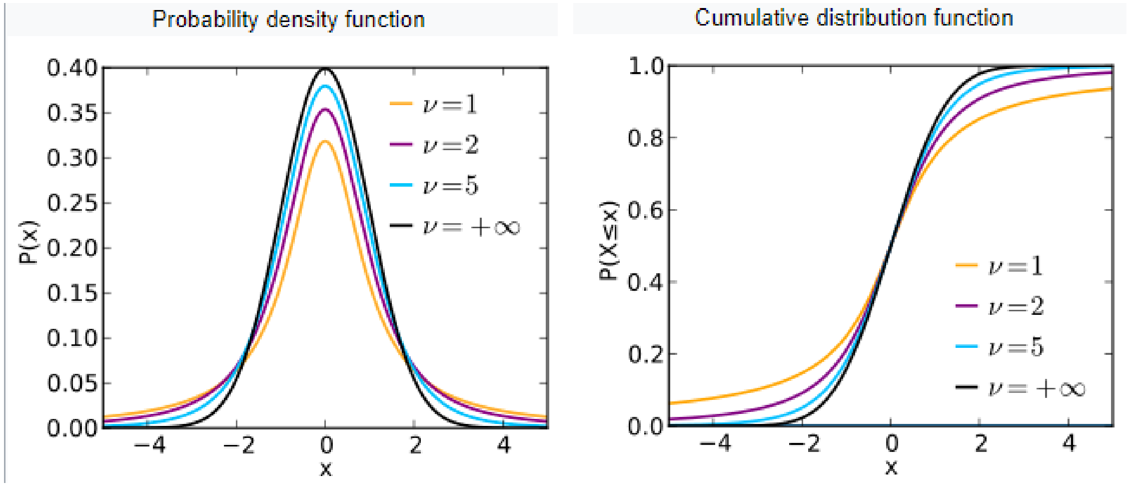
The distribution arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown.

It is the basis of the popular Student's t between two sample means (more on that soon!)

The pdf is given by:

$$f(x) = \frac{\Gamma((v+1)/2)}{\sqrt{v\pi}\Gamma(v/2)}(1+x^2/v)^{-(v+1)/2}$$

Where v is known as „degrees Of freedom”



Student’s t distribution

Exercise 4: Student’s t distribution

Display the Student's t distributions with 1,2,4 and 30 degrees of freedom and compare it to the normal distribution.

Distributions in R

We will use R to fit the distribution to some data.

Exercise 6:

Create 1000 random sampling from log-normal distribution.
Verify the values of the parameters of the distribution.

Exercise 7:

Use data on air quality available in R.
Use variable describing temperature in New York and fit its distribution assuming:
(1) normal distribution;
(2) log-normal distribution;
(3) gamma distribution.

Distributions in R

Distribution	Functions			
Discrete	**CDF value**	**PMS/PDF value**	**Inverse CDF - F^{-1}**	**Generating random samplings from a given disrtibution**
Binomial	pbinom	dbinom	qbinom	rbinom
Beta	pbeta	dbeta	qbeta	rbeta
Poisson	ppois	dpois	qpois	rpois
Geometric	pgeom	dgeom	qgeom	rgeom
Hypergeometric	phyper	dhyper	qhyper	rhyper
Negative Binomial	pnbinom	dnbinom	qnbinom	rnbinom
Continous				
Normal	pnorm	dnorm	qnorm	rnorm
Log Normal	plnorm	dlnorm	qlnorm	rlnorm
Gamma	pgamma	dgamma	qgamma	rgamma
Chi-Square	pchisq	dchisq	qchisq	rchisq
Student t	pt	dt	qt	rt
Cauchy	pcauchy	dcauchy	qcauchy	rcauchy
Exponential	pexp	dexp	qexp	rexp
F	pf	df	qf	rf

Bibliography

Christian Heumann, Michael Schomaker Shalabh „Introduction to Statistics and Data Analysis With Exercises, Solutions and Applications in R", Springer 2016: Chapters 1&8

Thank you for your attention

Time for practice!