02403 Introduction to mathematical Statistics

Lecture 1: Introduction, Summary statistics, Python, and Random variables

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

IntroStat team

Pernille Y. Nielsen

M S Khalid

Jan K. Møller

Nicolai S. Larsen

Peder Bacher

- 02402 Statistics (Polytechnical Foundation)
- 02323 Introduction to Statistics
- 02403 Introduction to Mathematical Statistics

Agenda

1. Practical course information
2. Introduction and Motivation
3. Descriptive Statistics
   - Percentiles and quantiles
4. Software: Python in Visual Studio Code
5. Random variables
6. Concrete discrete distributions

Overview

1. Practical course information
2. Introduction and Motivation
3. Descriptive Statistics
   - Percentiles and quantiles
4. Software: Python in Visual Studio Code
5. Random variables
6. Concrete discrete distributions

Practical course information

- Exam
  - June 26, 2025
  - 4-hour multiple choice
- Mandatory project
  - 1 project must be approved to participate in the exam.
  - Project available on our website
  - If you've already passed the project, you do not need to submit again.

Practical course information

- Generic agenda
  - Before teaching: Read relevant sections in the book
  - Lectures: 2 hours, curriculum of the day
  - Exercises: 2 hours
  - After teaching: Exercises, previous exam questions, and reading
- Teaching material
  - Available under *Agendas* on course website
  - Lecture slides and Python code might be updated shortly before each lecture (remember to refresh browser)

Practical Information

- Website: 02403.compute.dtu.dk
  - Agendas: Teaching plan, syllabus, exercises and solutions (English), slides, scripts, previous exams and solutions
  - Book
  - Project (&Deadline) - Submit via Learn
  - Podcasts of previous years' lectures (OBS: no recordings using Python)
- DTU Learn
  - Announcements
  - Projects - description and submission

Python

In addition to paper and pencil, we use Python to calculate statistical tasks (you have to do the project in Python).

Overview

1. Practical course information
2. Introduction and Motivation
3. Descriptive Statistics
    - Percentiles and quantiles
4. Software: Python in Visual Studio Code
5. Random variables
6. Concrete discrete distributions

Preamble

Statistics is a mathematical science of collecting, describing, analyzing, and interpreting data.
You will extract knowledge and learn from observed data.

Probability is a branch of mathematics that deals with the description and analysis of chance.
You will derive knowledge and learn from a theoretical model

The fields are difficult to separate and methods from both fields are used commonly together in engineering.

Et fælles mål: A common goal: Describe and understand random variation and uncertainties quantitatively!

History & Application of Statistics in Various Fields

There are many exciting research areas within both applied and theoretical statistics.

Statistics are used in many areas, for example:

- Clinical studies and epidemiology
- Production, planning and quality control
- Analysis of laboratory data and experiments
- Data Science and Artificial Intelligence (AI)
- Forecasting

Some historic notes

- Development of mathamatics of probability theory (dating back several of millenniums)
- Cinical trials (e.g. Janmes Lind, 1716-1794)
- Epidemiology (John Snow, 1813-1858)
- Ronald Fisher (1890-1962) is considered the founder of modern statistics

In Everyday Life

Statistics or elements from the subject appear in many places in everyday life, including:

- News
- Politics
- Advertising
- Sports
- Work

Statistics are often used as decision support! Statistics can be used to determine what should be investigated further.

Common Fallacies and Biases

Statistics can be counterintuitive, and our brains need to be trained in statistical thinking to avoid making a series of common fallacies. *Even well-educated, professional statisticians make simple mistakes*.

Some typical biases (systematic distortions) in statistics are:

- Survivorship bias: Ignoring failures, focusing on successes.
- Selection bias: Distorted results due to sample non-representativeness.
- OVB (Omitted-variable bias): Missing variables skew results and conclusions.

The Course's Overall Goals and Scope

The course should, among other things, help you to:

- Handle and analyze data appropriately
- Describe and understand random variation and uncertainties
- Think critically about statistical statements
- Understand the possibilities and limitations of statistics

The course should also prepare you for advanced courses in experimental design, time series analysis, quality control, probability theory, statistical modelling, data analysis, machine learning, and artificial intelligence. Limitations: (1) Small datasets, (2) Simplified scenarios lack real-world complexities and context.

The course content in broad terms

A large part of the course covers:

1. Formulation of models
2. Calculation of confidence intervals
3. Conducting hypothesis tests
4. determining whether various conditions are *statistically significant*

in different contexts and setups.

Probability theory will be our primary tool.

Basics of Statistics

Statistics can generally be divided into two parts:

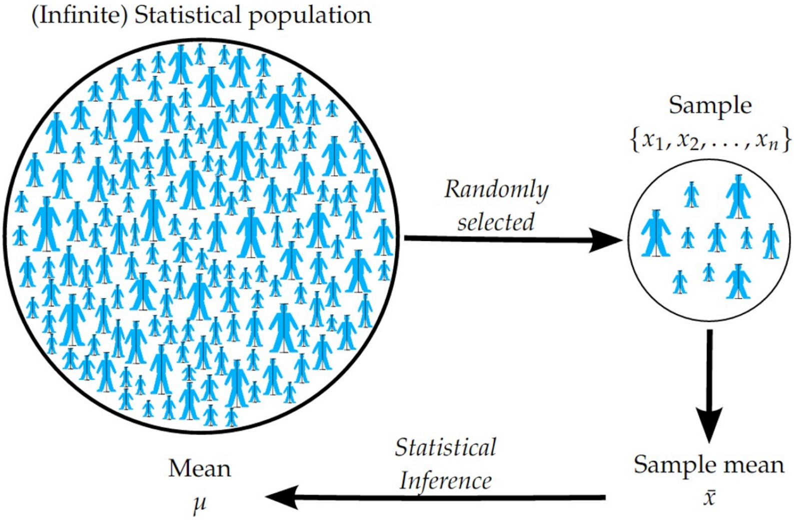
- Descriptive statistics
- Inferential statistics

Statistics typically involves analyzing a *sample* taken from a *population*.

From the sample, we generally make statements about the population.

It is therefore important that the sample is *representative* of the population. *For most of the course, we will simply assume that the samples are representative.*

Statistics


(Infinite) Statistical population
Sample
$\{x_1, x_2, \ldots, x_n\}$
Randomly selected
Sample mean
$\bar{x}$
Statistical Inference
Mean
μ

Overview

1. Practical course information
2. Introduction and Motivation
3. Descriptive Statistics
    - Percentiles and quantiles
4. Software: Python in Visual Studio Code
5. Random variables
6. Concrete discrete distributions

The General Setup

There is an underlying population from which a representative sample with n observations has been drawn.

The sample is usually represented by a vector

$$y = (y_1, y_2, \ldots, y_n).$$

The sorted sample is then

$$(y_{(1)}, y_{(2)}, \ldots, y_{(n)}),$$

where $y_{(1)}$ denotes the smallest observation and $y_{(n)}$ denotes the largest observation.

Summary statistics (Nøgletal)

We use *summary statistics* to summarize and describe data (random variables)

- **Average:** Measure of center / location
- **Median:** Measure of center / location
- **Variance:** Variation
- **Standard deviation:** Variation (same unit as data)
- **Coefficient of variation:** Variation in data (unit less)
- **Covariance:** (linear) interdependence
- **Correlation:** (linear) interdependence (unit less)
- **Quantiles:** For making statements about the data distribution

Summary statistics (Nøgletal)

Let $(y_{11},...,y_{1n})$ and $(y_{21},...,y_{2n})$ be samples

		Definition	Python
Average:	$\bar{y}_1$	$\frac{1}{n}\sum_{i=1}^{n} y_{1i}$	`np.mean(y1)`
Variance:	s_1^2	$\frac{1}{n-1}\sum_{i=1}^{n}(y_{1i}-\bar{y}_1)^2$	`np.var(y1, ddof = 1)`
Standard deviation:	s_1	$\sqrt{s^2}$	`np.std(y1, ddof = 1)`
Covariance:	s_{12}	$\frac{1}{n-1}\sum_{i=1}^{n}(y_{1i}-\bar{y}_1)(y_{2i}-\bar{y}_2)$	`np.cov(y1, y2)`
Correlation:	r_{12}	$\frac{s_{12}}{s_1 \cdot s_2}$	`np.corrcoef(y1, y2)`

Summary statistics

Let $y_{(1)} \leq y_{(2)} \leq ... \leq y_{(n)}$ be the sortered sample of $y_1, ..., y_n$.

		Definition	Python
Coefficient of variation:	V	$\frac{s}{\bar{y}} \cdot 100$	
Quantiles:	τ_p	$\frac{\left(y_{(np)}+y_{(np+1)}\right)}{2}$, $np = \lceil np \rceil$ $y_{(\lceil np \rceil)}$, $np \neq \lceil np \rceil$	`np.percentile(y,...)`[1]
First quartile	Q_1	$\tau_{0.25}$	
Median	$\tilde{y}$	$\tau_{0.50}$	`np.median(y)`
Third quartile	Q_3	$\tau_{0.75}$	
Inter Quartile Range	IQR	$Q_3 - Q_1$	

[1] use `np.percentile(y, probs = p, method = "averaged_inverted_cdf")` for the definition above

Example: Student heights

- **Sample:** Student heights in cm, $n = 5$.

$$(y_1, y_2, y_3, y_4, y_5) = (185, 184, 194, 180, 182)$$

- **Average:**

$$\bar{y} = \frac{1}{5}(185 + 184 + 194 + 180 + 182) = 185$$

- **Median:**
    - First order the data: $180, 182, 184, 185, 194$.
    - Then choose the third/middle number (n uneven): 184
- If a person with height 235 cm is added to the data:
    - *Average:* 193
    - *Median:* 184.5

Example on Dispersion/Spread: Student heights

- **Sample:** Student heights in cm, $n = 5$.

$$(y_1, y_2, y_3, y_4, y_5) = (185, 184, 194, 180, 182)$$

- **Variance:**

$$s^2 = \frac{1}{4}((185 - 185)^2 + (184 - 185)^2 + \cdots + (182 - 185)^2) = 29$$

- **Standard deviation:**

$$s = \sqrt{29} = 5.385$$

Percentiles and quantiles

The median is the value that divides the data into two halves. More generally, we may compute *percentiles*, e.g.:

- $0, 25, 50, 75, 100$ % percentiles and/or
- $0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$ % percentiles

Note:

- The median is the 50% percentile.
- The 25, 50, 75 % percentiles are often referred to as the first, second and third quartiles, and denoted $Q1$, $Q2$, and $Q3$, respectively.
- Inter Quartile Range (IQR): $Q3 - Q1$

Example: Student heights

- **Sample:** *Ordered* student heights in cm.

$$(y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}) = (180, 182, 184, 185, 194)$$

- **Lower quartile, Q1:**
  - Establish that $np = 1.25$, as $p = 0.25$ and $n = 5$.
  - The smallest integer larger than np is 2.
  - $Q1 = y_{(2)} = 182$.
- **Upper quartile, Q3:**
  - Establish that $np = 3.75$, as $p = 0.75$ and $n = 5$.
  - The smallest integer larger than np is 4.
  - $Q3 = q_{0.75} = y_{(\lceil 3.75 \rceil)} = y_{(4)} = 185$.
- **IQR:**
  - $Q3 - Q1 = 185 - 182 = 3$.

Sample covariance and correlation

Student	1	2	3	4	5	6	7	8	9	10
Heights (y_{1i})	168	161	167	179	184	166	198	187	191	179
Weights (y_{2i})	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9
$(y_{1i} - \bar{y}_1)$	-10	-17	-11	1	6	-12	20	9	13	1
$(y_{2i} - \bar{y}_2)$	-12.6	-19.8	-10	7.6	2.4	-14.7	24.5	13.3	8.6	0.8
$(y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)$	126.1	336.8	110.1	7.6	14.3	176.5	489.8	119.6	111.7	0.8

$$\begin{aligned} s_{12} &= \frac{1}{9}(126.1 + 336.8 + 110.1 + 7.6 + 14.3 + 176.5 + 489.8 \\ &\quad +119.6 + 111.7 + 0.8) \\ &= \frac{1}{9} \cdot 1493.3 \\ &= 165.9 \end{aligned}$$

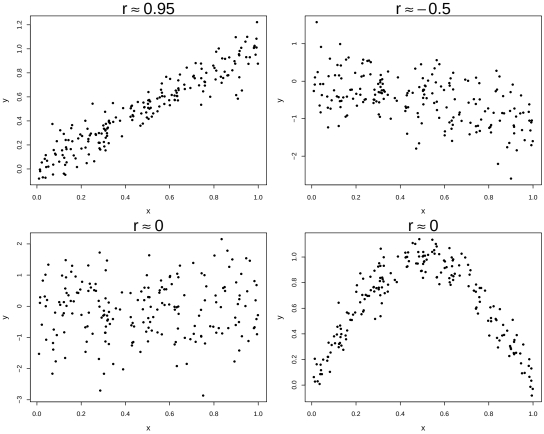
$$s_1 = 12.21, \quad \text{and} \quad s_2 = 14.07$$

$$r = \frac{165.9}{12.21 \cdot 14.07} = 0.97$$

Properties of Correlation Coefficient

- r is always between -1 and 1: $-1 \leq r \leq 1$.
- r measures the degree of linear relation between x and y.
- $r = \pm 1$ if and only if all points in the scatterplot are exactly on a line.
- $r > 0$ if and only if the general trend in the scatterplot is positive.
- $r < 0$ if and only if the general trend in the scatterplot is negative.

Correlation



Second order moment representation

Let $\boldsymbol{y}_i = \left[y_{i,1}, \ldots, y_{i,k}\right]$, the average vector is

$$\bar{\boldsymbol{y}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{y}_i.$$

The observed variance-covariance matrix is given by

$$\boldsymbol{S} = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T.$$

with $\boldsymbol{S}_{lm} = s_{lm}$. The pair $\bar{\boldsymbol{y}}$ and $\boldsymbol{S}$ is reffered to as the *second order moment* representation. The variance covariance matrix is often decomposed into standard deviation and correlation

$$\boldsymbol{S} = \hat{\boldsymbol{\sigma}} \boldsymbol{R} \hat{\boldsymbol{\sigma}},$$

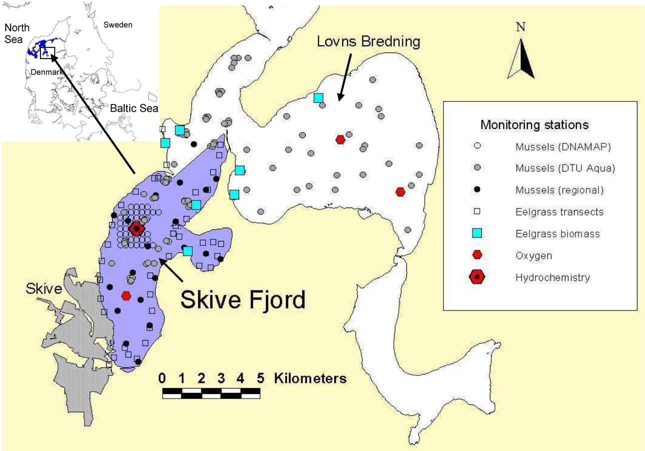
where $\hat{\boldsymbol{\sigma}}$ is a $k \times k$ matrix with $\hat{\boldsymbol{\sigma}}_{ii} = \sqrt{\boldsymbol{S}_{ii}}$, $\hat{\boldsymbol{\sigma}}_{ij} = 0$ for $i \neq j$, and $\boldsymbol{R}$ is the collection of all pairwise correlations

Figures/Tables

- Quantitative data
  - Scatter plot (xy plot), related to correlation
  - Histogram, related to density
  - Cumulative distribution, related to cumulative distribution function
  - Box plot, summary of distribution
- Qualitative/Count data
  - Bar chart
  - Pie chart

Visualization of data is important!
We can make different figures and tables in Python

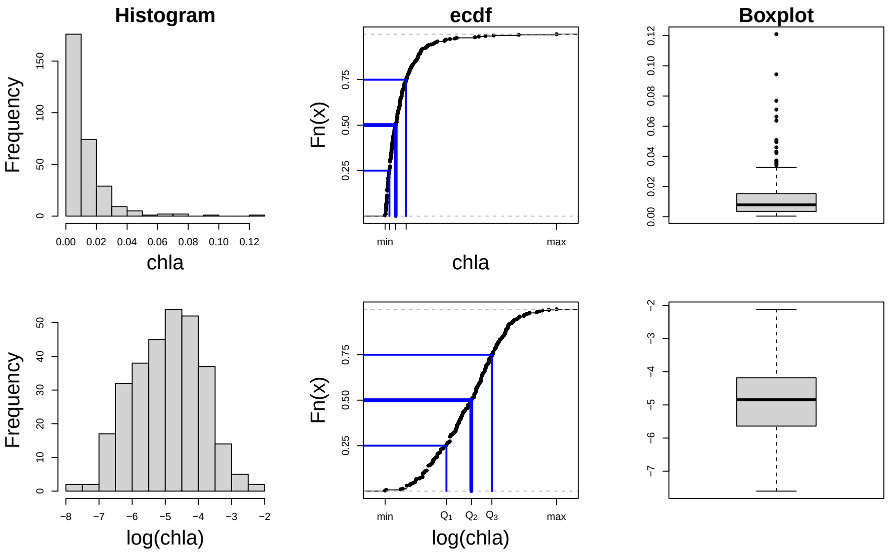
Example: Skive fjord



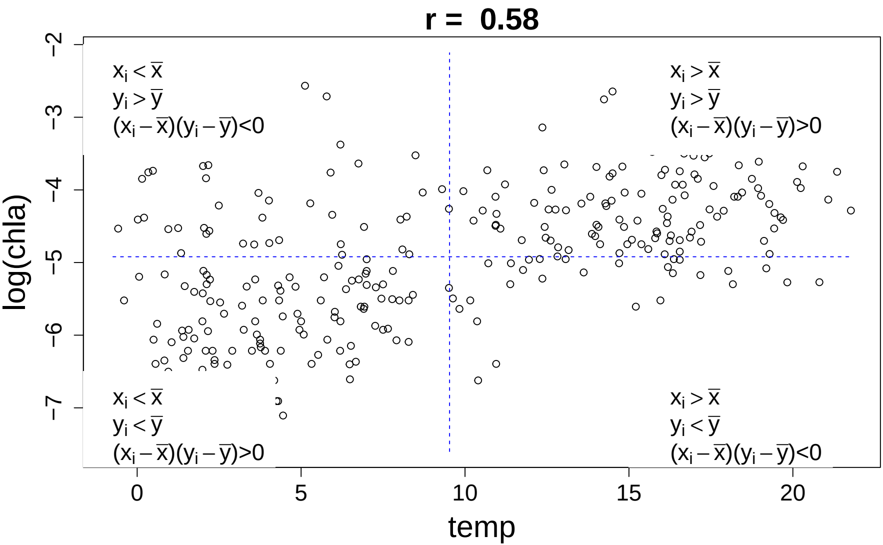
Eksempel: Skive fjord, data

The data set contain average monthly observations of a number of different variables, today we will consider chloryphyll measurements.

Example: Clorophyl distribution



Example: Scatterplot



Overview

1. Practical course information
2. Introduction and Motivation
3. Descriptive Statistics
   - Percentiles and quantiles
4. Software: Python in Visual Studio Code
5. Random variables
6. Concrete discrete distributions

Software: Python in Visual Studio Code

- Python: An interpreted, interactive, object-oriented programming language. It provides high-level data structures such as list and associative arrays (dictionaries), dynamic typing and dynamic binding, modules, classes, exceptions, automatic memory management, etc.
- Python: Free, open source, works on all platforms.
- Many Python "libraries" for all kinds of data analysis.
- Introduced in the book & ingeral part of the course.
- Learning by doing.
- Visual Studio Code as code editor using Microsoft Python extension

Software: Python in Visual Studio Code

- We use VS Code to edit and run our Python programs.
- We primarily use Jupyter notebooks (.ipynb).
- We expect that you have installed both Python and VS Code and that you can create a Jupyter notebook.
- Help with installation: https://pythonsupport.dtu.dk/

Python abbreviations in the course

In this course, we use a number of Python "libraries" and apply the following abbreviations:

- import numpy as np
- import matplotlib.pyplot as plt
- import pandas as pd
- import scipy.stats as stats
- import statsmodels.api as sm
- import statsmodels.formula.api as smf
- import statsmodels.stats.power as smp
- import statsmodels.stats.proportion as smprop

Python

- Go to today's Python notebook in VS Code

Overview

1. Practical course information
2. Introduction and Motivation
3. Descriptive Statistics
   - Percentiles and quantiles
4. Software: Python in Visual Studio Code
5. Random variables
6. Concrete discrete distributions

Random variables

A random variable represents a value of an outcome *before* the corresponding *experiment* is carried out.

- A throw of a dice.
- The number of six'es in ten dice throws.
- Fuel consumption of a car.
- Measurement of glucose level in blood sample.
- ...

Discrete and continuous random variables

- We distinguish between *discrete* and *continuous* random variables.
- Discrete (countable sample space):
  - Number of people in this room who wear glasses.
  - Number of planes departing from CPH within the next hour.
- Continuous (uncountable sample space):
  - Wind speed measurement.
  - Transport time to DTU.

Random variable

Before the experiment is carried out, we have a random variable

$$Y \text{ (or } Y_1, \ldots, Y_n)$$

indicated with capital letters.
Formally Y is a function that assign probabilities to subsets of possible outcomes, e.g. if Y is the number rolled with a fair dice then $P(Y = 1) = \frac{1}{6}$ and $P(Y \in \{1, 2\}) = \frac{2}{6}$.

After the experiment is carried out, we have a *realization* or *observation*

$$y \text{ (or } y_1, \ldots, y_n)$$

indicated with lowercase letters. y is a number (i.e. NOT a random variable), e.g. we roll 2 with a fair dice.

Discrete distributions

For discrete distributions

- The sample space is countable, e.g. $Y \in \{0, 1, ...\}$, i.e. $Y \in \mathbb{N}_0$
- The sample space is countable and finite $Y \in \{0, 1, ..., N\}$, $N < \infty$
- We have a model that describe the probability of each outcome, i.e. $P(Y = i)$

Example: For a roll with a fair dice then $Y \in \{1,2,3,4,5,6\}$, and $P(Y = i) = \frac{1}{6}$ for $i \in \{1,2,3,4,5,6\}$, and $P(Y = i) = 0$ for $i \notin \{1,2,3,4,5,6\}$.

Density function, discrete random variable: Definition 2.6

The probability density (mass) function, (pdf/pmf) of a discrete random variable:

Definition

$$f(y) = P(Y = y)$$

Describes the probability that Y takes the value y when the experiment is carried out. The density function of a discrete random variable satisfies two properties:

$$f(y) \geq 0 \text{ for all } y \quad \text{and} \quad \sum_{\text{all } y} f(y) = 1$$

The random variable may also be described by the *distribution function* (cumulative distribution function, cdf)

$$F(y) = P(Y \leq y) = \sum_{j \text{ where } y_j \leq y} f(y_j)$$

Mean and variance of descrete random variable

The mean and varaince of a discrete random variable are defined as

Definition

$$\begin{aligned}\mu =& E(Y) = \sum_{\text{all } y} y f(y)\\ \sigma^2 =& \text{Var}(Y) = \sum_{\text{all } y} (y-\mu)^2 f(y)\end{aligned}$$

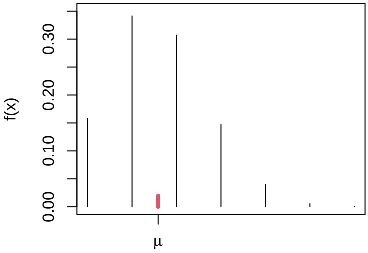
we can interpret these as the average and empirical variance we would get as $n \to \infty$. Formally if y_i is a ralization of the random variable Y_i with pdf $f(y)$ then

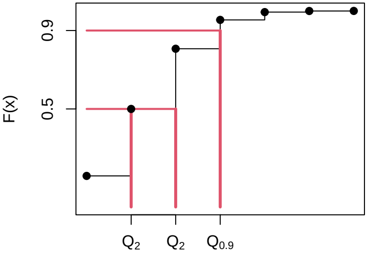
$$\lim_{n\to\infty} \hat{\mu} = \lim_{n\to\infty} \bar{y} = \mu$$

$$\lim_{n\to\infty} \hat{\sigma^2} = \lim_{n\to\infty} s^2 = \sigma^2$$

Summary statistics (Empirical and model based)

	Empirical	Discrete random variable
Mean	$\bar{y} = \sum y_i \frac{1}{n}$	$\mu = \sum y_i f(y_i)$
Variance	$s^2 = \sum (y_i - \bar{y})^2 \frac{1}{n-1}$	$\sigma^2 = \sum (y_i - \mu)^2 f(y_i)$
Median	$y_{(\lceil n/2 \rceil)}$[1]	"$F^{-1}(0.5)$"[2]
Quantile	Q_τ[1]	"$F^{-1}(\tau)$"





[1] see slide 23 for precise definition

[2] More precisely: x s.t. $P(Y \leq y) \geq 0.5$ and $P(Y \geq y) \geq 0.5$

Overview

1. Practical course information
2. Introduction and Motivation
3. Descriptive Statistics
   - Percentiles and quantiles
4. Software: Python in Visual Studio Code
5. Random variables
6. Concrete discrete distributions

Discrete distributions used in this course

We will consider the discrete distributions

- The Binomial distribution $B(n,p)$
- The hypergeometric distribution $H(n,a,N)$
- The Poisson distribution $P(\lambda)$

Each sutiable for different situations.

Discrete distributions used in this course

Distribution	f(y)	μ	σ^2	Typical application
$Y \sim B(n,p)$	$\binom{n}{y} p^y (1-p)^{n-y}$	np	$np(1-p)$	Flip a coin n-times (succes prob p).
$Y \sim H(n,a,N)$	$\frac{\binom{a}{y}\binom{N-a}{n-y}}{\binom{N}{n}}$	$n\frac{a}{N}$	$n\frac{a}{N}\frac{(N-a)}{N}\frac{N-n}{N-1}$	Number of white balls drawn from an urn with N balls and a white balls.
$Y \sim P(\lambda)$	$\frac{\lambda^y}{y!}e^{-\lambda}$	λ	λ	Number of arivals per hour when average number of arivals per hour is λ.

- There exist a number of other discrete distributions, fit for different porpuses.

Distributions in Python

We use `SciPy.stats` for most distributions in this course (but also `NumPy.random.choice` for dice roll etc.).

Documentation for `SciPy.stats` can be found online.

General 'methods' for most distributions are:

`scipy.stats`	`.binom/.hypergeom/.poisson`
`.rvs`	'random variates' (simulate random numbers)
`.pmf`	'probability mass function' (pmf/pdf/density function)
`.cdf`	'cumulative distribution function' (distribution function)
`.ppf`	'percent point function' (invers cdf / quantile function)
`.mean /.var /.std`	'mean'/'variance'/'standard deviation'

Example

The card game Casino is played with an ordinary deck of cards, i.e. 52 cards, 13 of each colour (clubs, hearts, diamonds and spade), and 4 of each value (e.g. 4 aces). The game starts by placing 4 cards face up on the table.[1]

- What is the probability that at least 3 of the 4 opening cards are aces?
- What is the probability that all the 4 opening cards are clubs in exactly 1 out of 4 games?

[1] 2021 June exam

Example

Two students are counting the number of cars passing by on different stretches of road. They assume that the number of cars passing by in specific time intervals follow Poisson distributions. On the first road (road 1) they assume that the expected number of cars passing by is $\lambda_1 = 10/\text{hour}$, while on the second road (road 2) they assume that the expected number of cars passing by is $\lambda_2 = 15/\text{hour}$. [2]
Now they define two random variables:

- X_1: number of cars passing by on road 1 in 15 minutes
- X_2: number of cars passing by on road 2 in 10 minutes.

You can assume that X_1 and X_2 are independent.

- What is the probability $P(X_1 = 10)$?
- What is $\frac{E[X_1]}{E[X_2]}$?
- What is the probability that the time between two cars passing by is greater than 2 minutes on road 2?

[2] 2024 June

Summary from today:

1. Practical course information
2. Introduction and Motivation
3. Descriptive Statistics
    - Percentiles and quantiles
4. Software: Python in Visual Studio Code
5. Random variables
6. Concrete discrete distributions

Exercises

- Exercises start at 10:15
- You can find the exercises on the course website's "Agenda" (the exercises, like the book, are only available in English)
    - Start by opening VS Code and creating a new notebook (check that it works by calculating 2+2)
    - Then proceed with the day's exercises (both with and without Python)
    - It's also a good opportunity to discuss the curriculum and ask questions to the teaching assistants
- Go to website 02403.compute.dtu.dk for locations of the exercises and names of TAs
- Please start the project already today (at least do the first question)

Next lecture:

- Probability II: Book Chapter Two.