Practice sheet 3: Estimation/Confidence Intervals

Exercise 1

The dataset in `results.csv` includes `43,170` results of international football matches starting from the very first official match in 1872 up to 2019. The matches range from FIFA World Cup to FIFI Wild Cup to regular friendly matches. The matches are strictly men's full internationals and the data does not include Olympic Games or matches where at least one of the teams was the nation's B-team, U-23 or a league select team.

`results.csv` includes the following columns:

- `date` - date of the match
- `home_team` - the name of the home team
- `away_team` - the name of the away team
- `home_score` - full-time home team score including extra time, not including penalty-shootouts
- `away_score` - full-time away team score including extra time, not including penalty-shootouts
- `tournament` - the name of the tournament
- `city` - the name of the city/town/administrative unit where the match was played
- `country` - the name of the country where the match was played
- `neutral` - TRUE/FALSE column indicating whether the match was played at a neutral venue

1. Import the dataset into `Python`
2. Create a column representing the type of the result of the match: win/lose/draw
3. Estimate using the dataset the probability to win/lose/draw a soccer match. Estimate the corresponding 95% Confidence Intervals.
4. Provide the same estimation as in 3. by countries (Egypt, for example), by some types of the tournament, by home/away match. Provide each case with a graphical representation in each case.

Assignement 2 (Part I)

Analyze the probability of winning/losing of the countries in the different tournaments. Study this probability in terms of several factors (playing at home/away). Compare confidence intervals and draw the corresponding graphs.

Exercise 2

The dataset in `covid_data.csv` includes the records of two years 2020 and 2021 in the countries affected by the COVID-19 pandemic.

`covid_data.csv` includes the following columns:

- `date` the date
- `iso3c` the country code
- `country` country name
- `income` Income group according WB classification
- `region` Geographical region according WB classification
- `continent` the name of the continent
- `dcases` the reported daily cases
- `ddeaths` the reported daily deaths
- `population` population size of the country in 2019
- `weekdays` weekdays
- `month` months

1. Compute the 95% Confidence intervals of the daily average reported cases and deaths by day in Egypt in 2021. Make a plot representing these Confidence intervals.
2. Compute the 95% Confidence intervals of the daily average reported cases and deaths by day in Africa in 2021. Make a plot representing these Confidence intervals.

Assignement 2 (Part II)

Analyze the daily reported number of cases and deaths, the case fatality rate (the ratio between deaths and cases) from COVID-19. Compare between 2020 and 2021. Compare between regions, income, and continents.

Exercise 3

A psychological study was conducted to compare the reaction of men and women to a stimulus. Independent random samples of 55 men and 65 women were employed in this experiment. The result is shown as follows:

- Men: $\overline{y} = 4.5$ seconds and $s^2 = .45$ seconds
- Women: $\overline{y} = 3.1$ seconds and $s^2 = .33$ seconds

1. Write a Python code computing the 95% Confidence intervals of the Men and Women averages.
2. Make a plot comparing both confidence intervals.
3. What's your conclusion?

Exercise 4

Let's consider the following Python code

[5]:
```
import scipy
import numpy as np
from scipy.stats import norm,t
```

[9]:
```
t.ppf(.975,19)
```

[9]:
```
2.093024054408263
```

[12]:
```
t.ppf(.995,19)
```

[12]:
```
2.860934606449914
```

[7]:
```
norm.ppf(.975,0,1)
```

[7]:
```
1.959963984540054
```

[8]:
```
norm.ppf(.995,0,1)
```

[8]:
```
2.5758293035489004
```

1. Which of the following 95% CI of the mean computed from the same random sample with size 20 comes from the t-distribution:

```
(27.442969558309777, 30.33435989685853)
```

and

```
(27.398576408603965, 30.37875304656434)
```

2. Deduce the sample mean of the random sample from the previous 95% Confidence intervals
3. Deduce the Margin of error of each Confidence interval
4. Deduce the sample variance of the random sample
5. Deduce the population variance of the random sample
6. Compute now the 99% Confidence intervals of the mean using the methods normal distribution (known population variance) and (unknown population variance) t-distribution