DATA SCIENCE FOR ECONOMISTS

ECON 220 LAB

Jafet Baca-Obando

Week 10, Central Limit Theorem – 10/31/2025

Outline

01 Distribution of a sample mean

02 Central limit theorem

03 Distribution of a sample proportion

Announcement

- Extra credit earned in the Lab can only be applied to the Lab portion of your grade, which is worth 25% of your overall grade in ECON 220.
- The maximum score for the entire Lab component is capped at 25% of your final grade, even if extra credit would otherwise raise it higher.
- In short: Lab extra credit can help you get up to the full 25% for the lab, but it cannot raise your score in the theory component.

Importing required libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats
import os
path = os.getcwd()
```

✓ 3.7s Python

Load the dataset

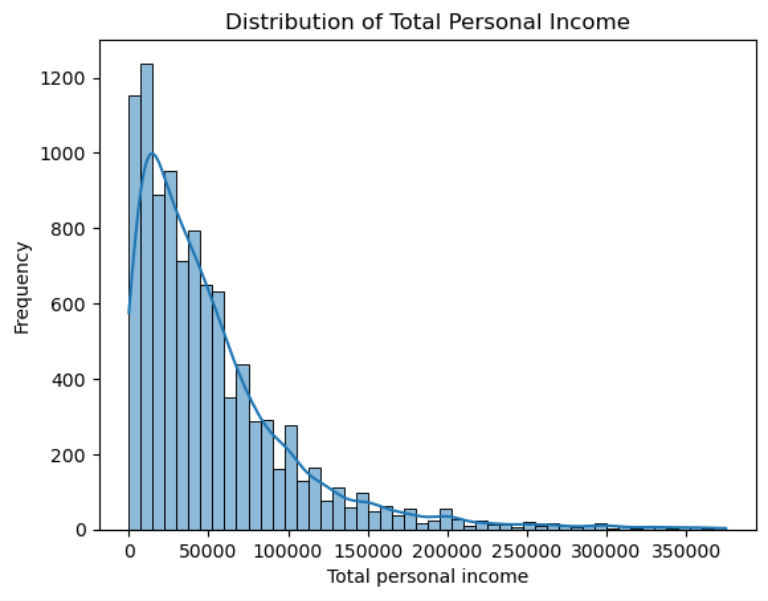
```
data = pd.read_csv("income.csv")
data.head(10)
```

✓ 0.0s Open 'data' in Data Wrangler Python

	# AGE	# INCTOT
0	39	89500
1	61	32000
2	36	33000
3	34	10000
4	46	35000
5	88	39000
6	58	241800
7	20	12000
8	69	15000
9	47	150000

- Our dataset: a sample of 10,000 individuals from 2023 IPUMS data.

Sampling distribution of the sample mean



- Distribution of personal income: right skewed, non-normal

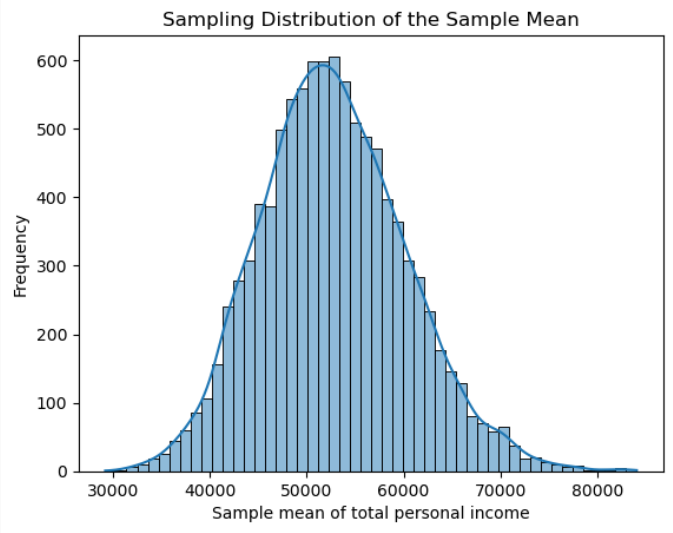
Sampling distribution of the sample mean

- Let's run a simulation. Instead of looking at the whole population, let's look at **sample means**.
  - Take a random sample of n = 50 people from the income data.
  - Calculate the mean income of that sample.
  - Repeat this process 10,000 times.
- This gives us a new dataset: a list of 10,000 sample means.

```python
N = 10000  # Number of samples
n = 50     # Sample size
sample_means = [] # List to store sample means
for _ in range(N):
    sample = data['INCTOT'].sample(n, replace=True)
    sample_means.append(sample.mean())
```

✓ 2.0s Python

The result: a normal distribution



- This is what we call the **Central Limit Theorem**.
- Mean of this sample: 52758.6. Population mean: 52635.0046

Central Limit Theorem

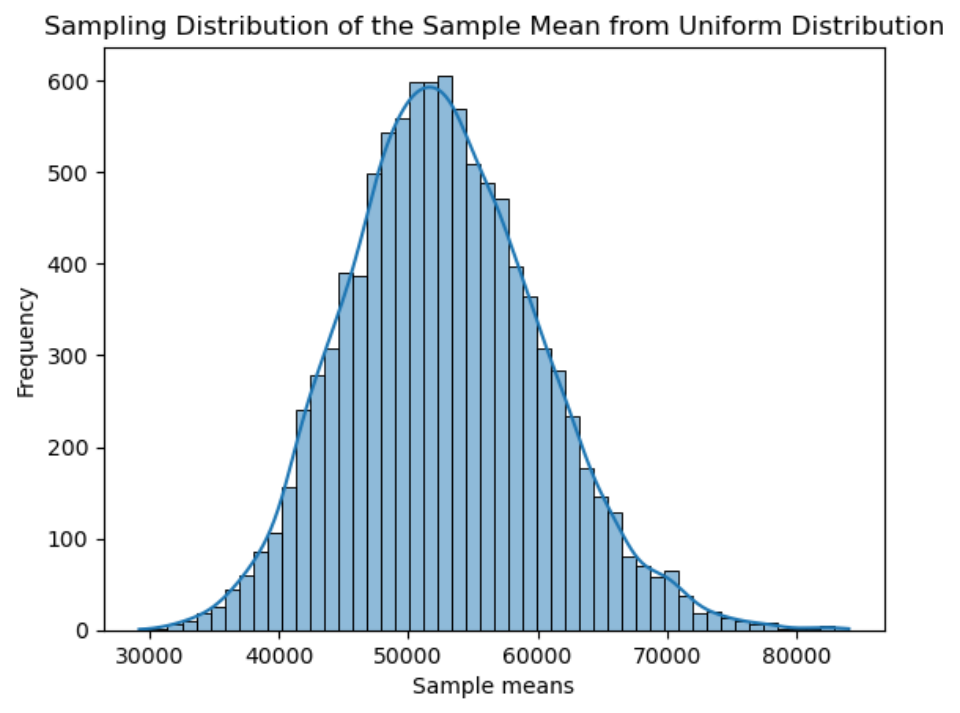
- The sampling distribution of the sample mean ($\bar{X}$) will **approach a normal distribution** as the sample size (N) increases, regardless of the population's original distribution.
- Rule of Thumb: A sample size of $N \geq 30$ is usually large enough for the CLT to apply.

Testing the CLT with the uniform distribution

```python
N = 10000  # Number of samples
n = 50     # Sample size
sample_means = [] # List to store sample means
for _ in range(N):
    s = np.random.choice(unif, size=n, replace=True)
    sample_means.append(s.mean())
```

✓ 0.2s Python

Testing the CLT with the uniform distribution



Sampling distribution for a sample proportion

- The CLT doesn't just apply to means—it also applies to **proportions**.
- Example: The unemployment rate in a population is 3.8% ($p = 0.038$). The exact probability of finding 3 or fewer unemployed people in a sample of 100 is given by the Binomial distribution.

```
stats.binom.cdf(3, 100, 0.038)
```

✓ 0.0s Python

```
0.4703111414330726
```

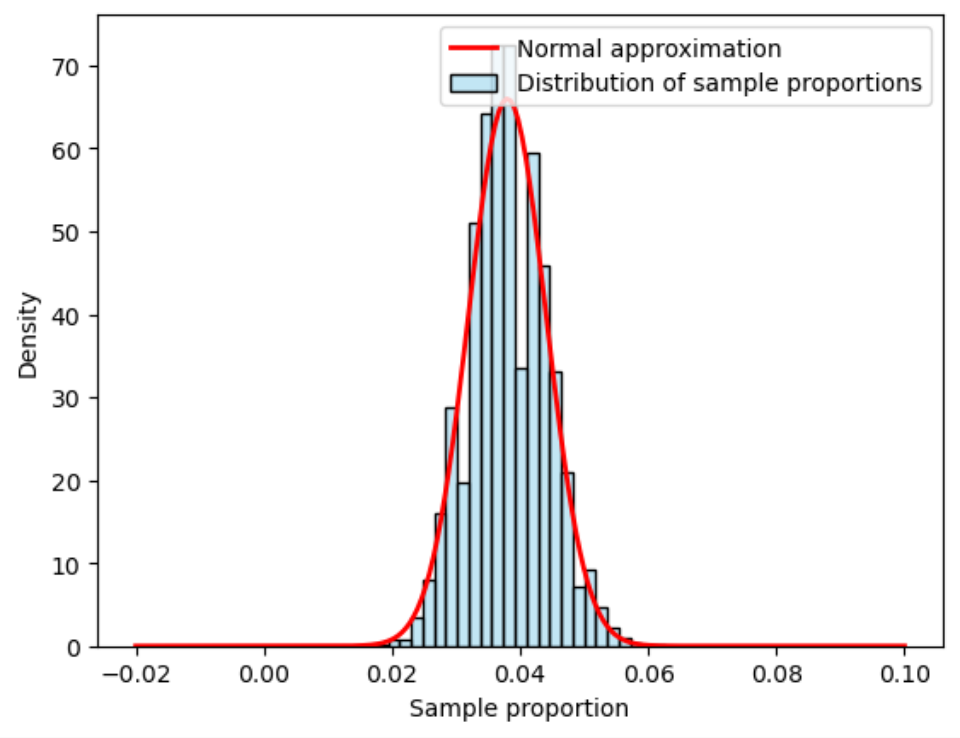
The CLT for proportions

- Instead of the Binomial, the CLT tells us we can approximate the sampling distribution of a proportion with a normal distribution with mean p and variance $\frac{p(1-p)}{n}$.

```python
N = 100000  # Number of samples
n = 1000     # Sample size
p = 0.038  # Success probability
sample_proportions = [] # List to store sample proportions
for _ in range(N):
    sample = stats.binom.rvs(n=n, p=p)
    proportion = sample/n
    sample_proportions.append(proportion)
```

✓ 5.7s Python

The CLT for proportions



Recap

- The CLT is a fundamental theorem that allows us to make inferences about a population from a sample.
- Even if a population is not normal, the distribution of its sample means will be approximately normal.
- This principle holds true for both sample means and sample proportions.
- A sample size of at least 30 is the general rule of thumb for the CLT to apply.
- Key to conducting hypothesis testing and building confidence intervals.

To-do list

- **Complete Data Exercise 6**
  - Upload Jupyter notebook (.ipynb file) and HTML file on **November 2**
- **Complete Data Exercise 8**
  - Upload Jupyter notebook (.ipynb file) and HTML file on **November 9**