CDT502: Data analytics and decision-making

Lecture: DDDM in Python

MScCT+ Program Core Courses

Dr. ZHI Shaohua
Friday, 25 April 2025

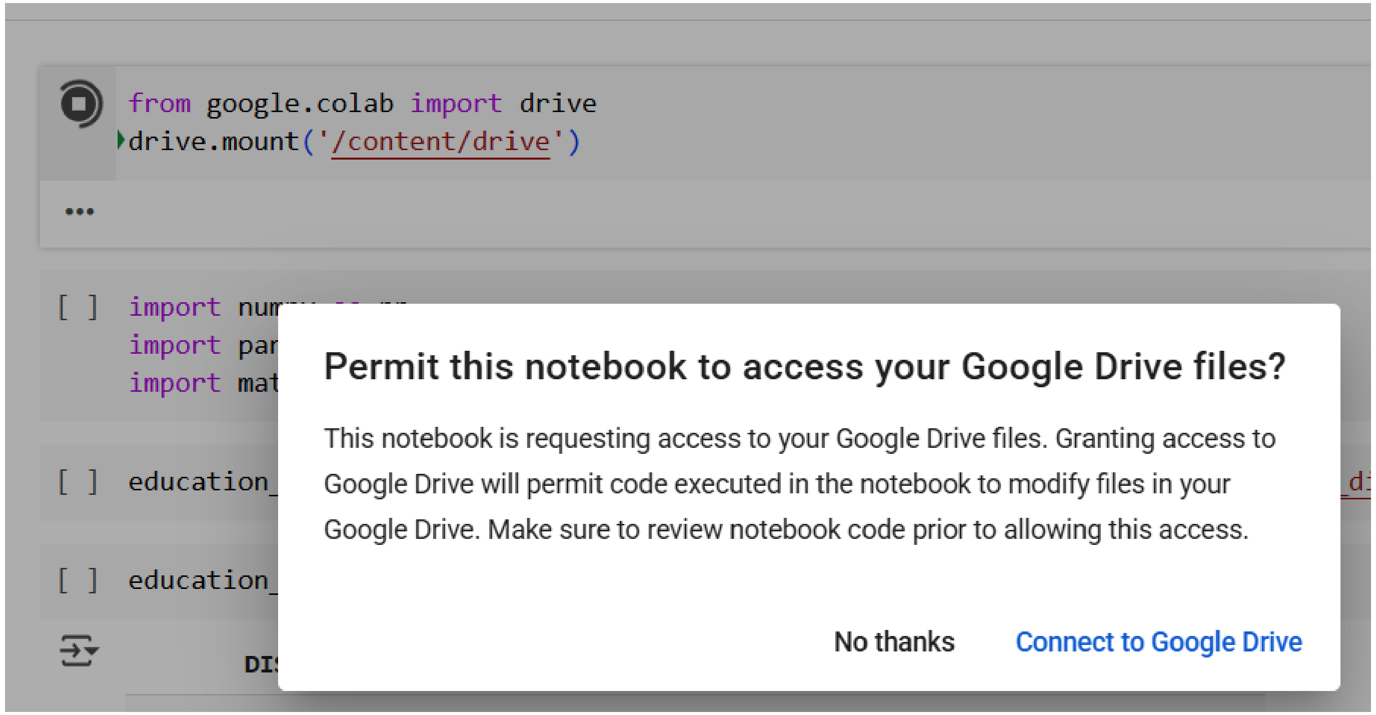
Content: Introduction to statistics

- Descriptive statistics with Python
- Probability distribution
- Sampling distribution
- Confidence Intervals
- Hypothesis Testing

Source Code

Descriptive statistics with Python

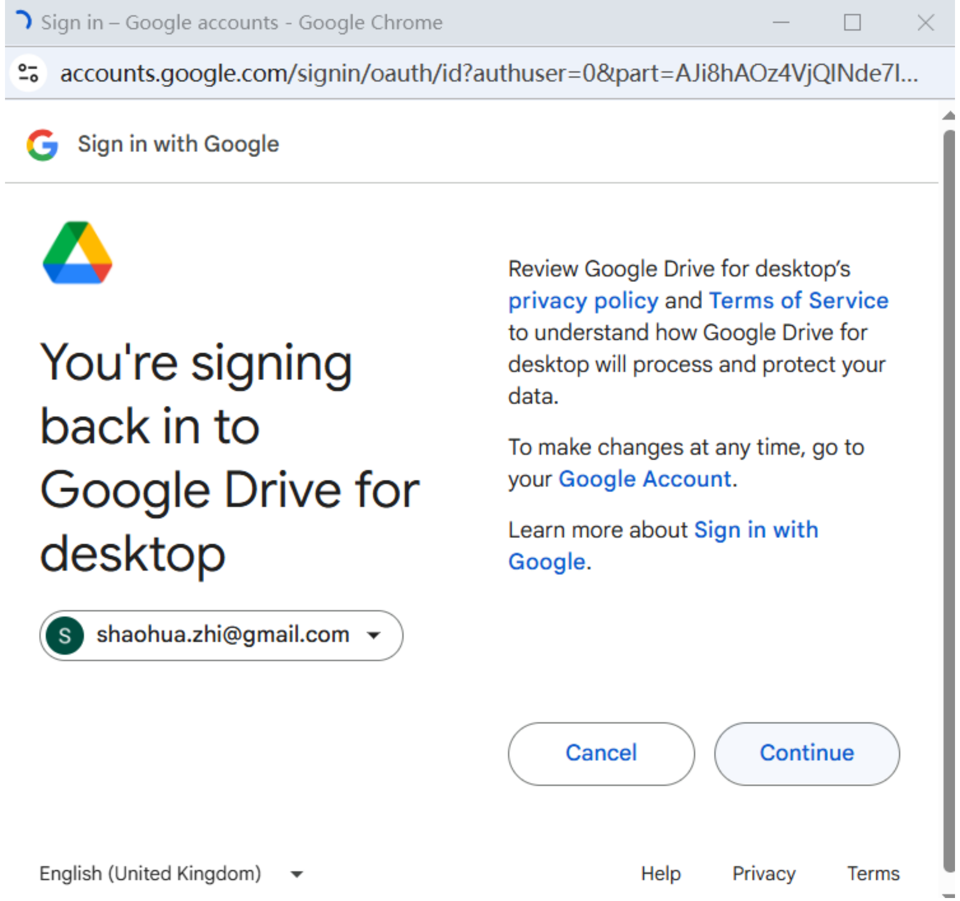
```
from google.colab import drive

drive.mount('/content/drive')
```



```
from google.colab import drive
drive.mount('/content/drive')
...
[ ] import num
    import par
    import mat
[ ] education_
[ ] education_
    DI
```

Permit this notebook to access your Google Drive files?

This notebook is requesting access to your Google Drive files. Granting access to Google Drive will permit code executed in the notebook to modify files in your Google Drive. Make sure to review notebook code prior to allowing this access.



Sign in – Google accounts - Google Chrome

accounts.google.com/signin/oauth/id?authuser=0&part=AJi8hAOz4VjQINde7I...

Sign in with Google

You're signing back in to Google Drive for desktop

shaohua.zhi@gmail.com

Review Google Drive for desktop's privacy policy and Terms of Service to understand how Google Drive for desktop will process and protect your data.

To make changes at any time, go to your Google Account.

Learn more about Sign in with Google.

Descriptive statistics with Python

- Import packages and libraries
- Explore the data

Functions for stats

.read_csv() – read the .csv file

.head() – get a quick overview of the dataset

.describe() – compute descriptive stats

float:

count
mean
std
min
25%
50%
75%
max

string:

count: Number of non-NA/null observations
unique: Number of unique values
top: The most common value (the mode)
freq: The frequency of the most common value

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html

Descriptive statistics with Python

Describe() for a numeric column

count

mean

std

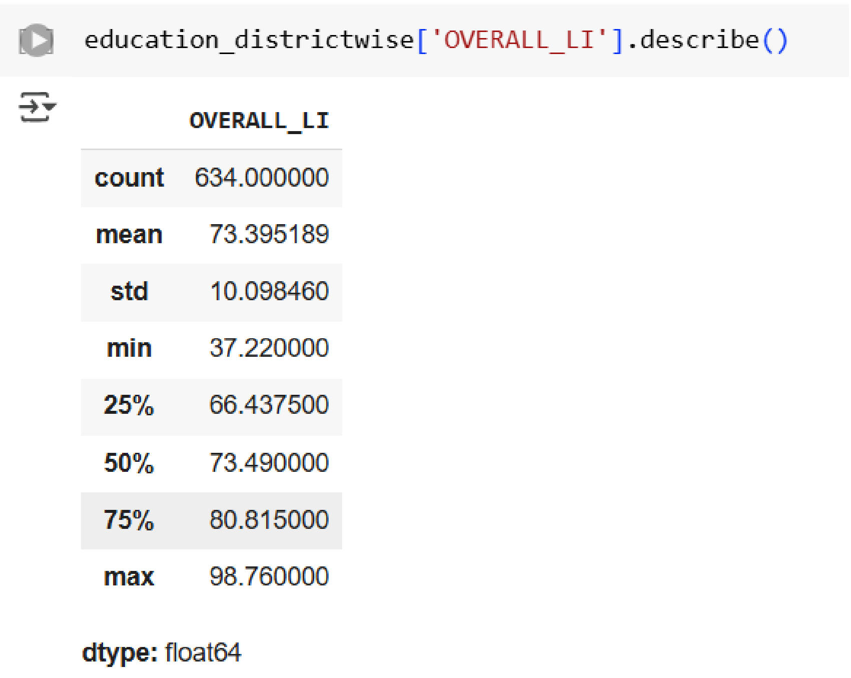
min

25%

50%

75%

max

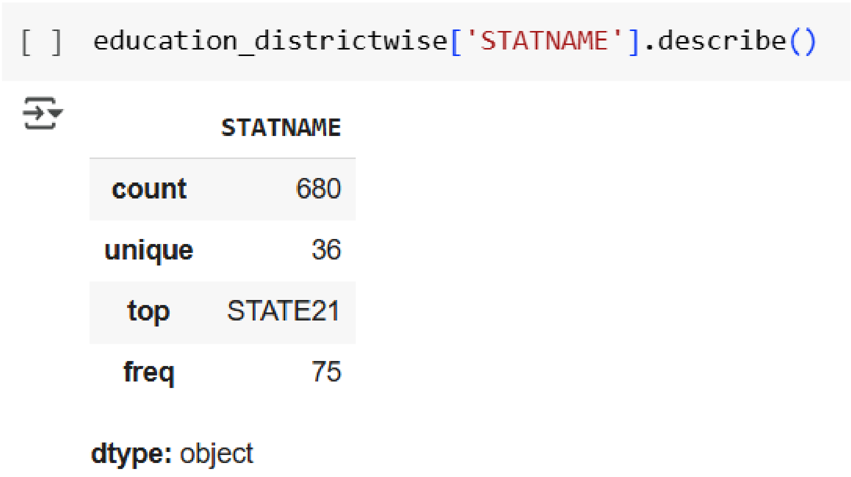


```python
education_districtwise['OVERALL_LI'].describe()
```

```
       OVERALL_LI
count  634.000000
mean    73.395189
std     10.098460
min     37.220000
25%     66.437500
50%     73.490000
75%     80.815000
max     98.760000

dtype: float64
```

Describe() for a categorical column

- count: Number of non-NA/null observations
- unique: Number of unique values
- top: The most common value (the mode)
- freq: The frequency of the most common value



```python
[ ]  education_districtwise['STATNAME'].describe()
```

```
          STATNAME
count          680
unique          36
top        STATE21
freq            75

dtype: object
```

Descriptive statistics with Python

- **Functions for stats**

.mean()

.median()

.min()

.max()

.mode()

.var()

.std()

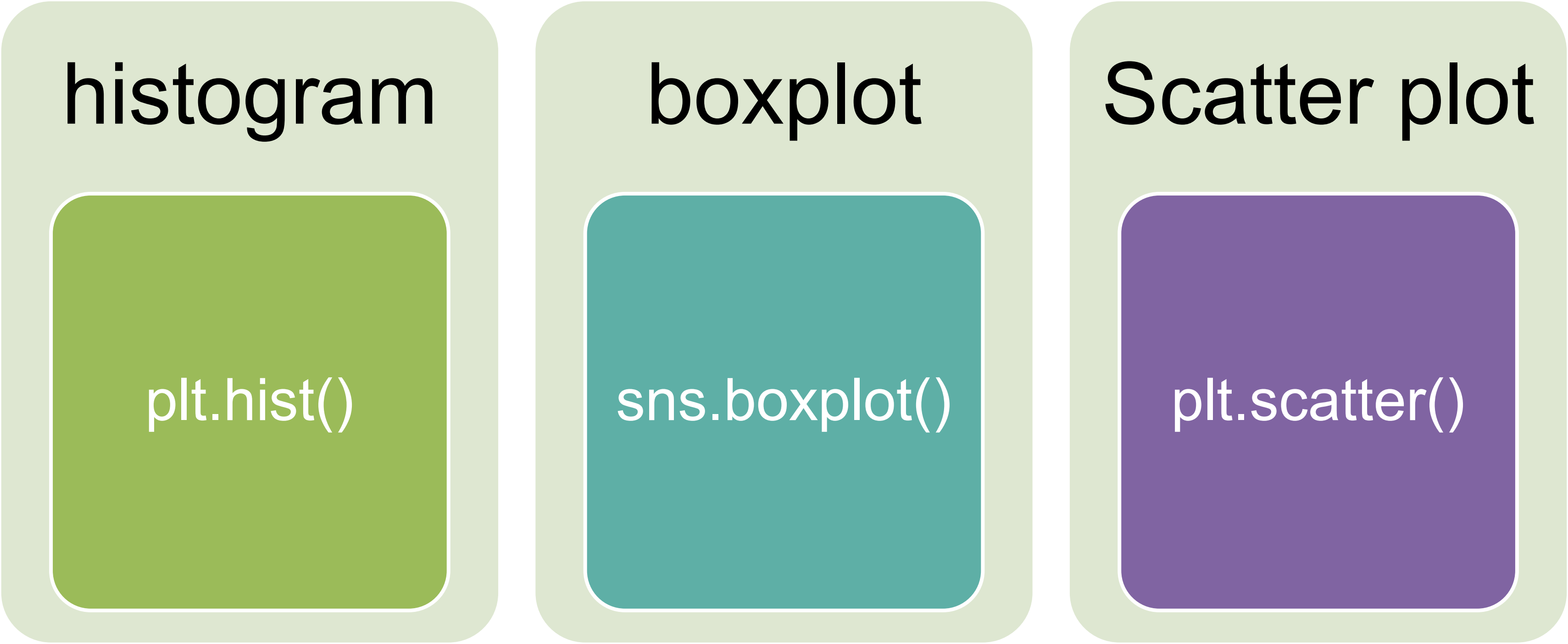
Q1, Q3, IQR

.quantile()

Descriptive statistics with Python

- **Visualization: Matplotlib, Seaborn**
  - **import library: seaborn**

```
import matplotlib.pyplot as plt
import seaborn as sns
```



histogram

plt.hist()

boxplot

sns.boxplot()

Scatter plot

plt.scatter()

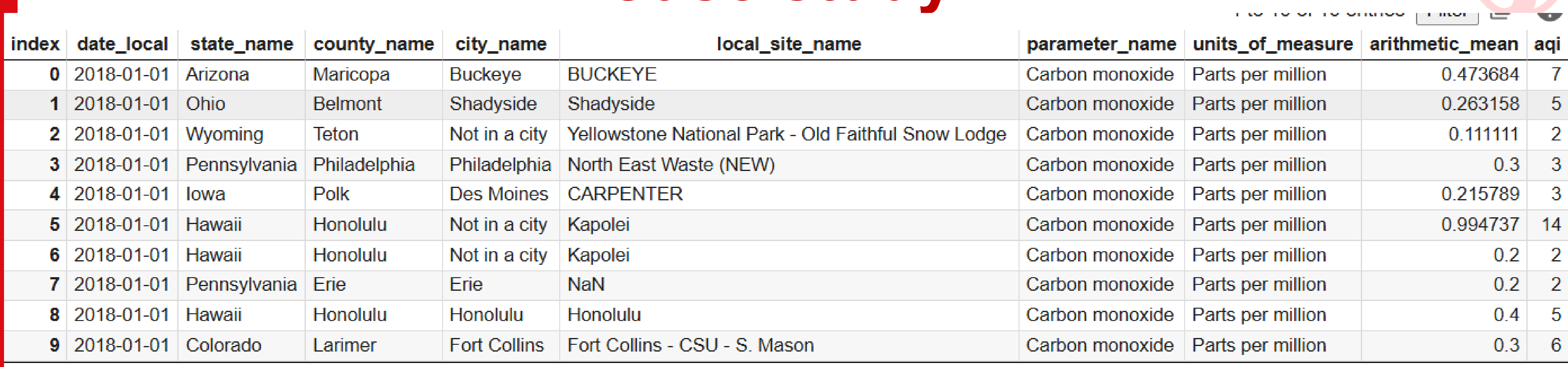
Case study

In this case, you are a member of an analytics team for the United States Environmental Protection Agency (EPA). You are assigned to analyze data on air quality with respect to carbon monoxide, a major air pollutant.

The data includes information from more than 200 sites, identified by state, county, city, and local site names. You will use Python functions to gather statistics about air quality, then share insights with stakeholders.

`c4_epa_air_quality.csv`

Case study



index	date_local	state_name	county_name	city_name	local_site_name	parameter_name	units_of_measure	arithmetic_mean	aqi
0	2018-01-01	Arizona	Maricopa	Buckeye	BUCKEYE	Carbon monoxide	Parts per million	0.473684	7
1	2018-01-01	Ohio	Belmont	Shadyside	Shadyside	Carbon monoxide	Parts per million	0.263158	5
2	2018-01-01	Wyoming	Teton	Not in a city	Yellowstone National Park - Old Faithful Snow Lodge	Carbon monoxide	Parts per million	0.111111	2
3	2018-01-01	Pennsylvania	Philadelphia	Philadelphia	North East Waste (NEW)	Carbon monoxide	Parts per million	0.3	3
4	2018-01-01	Iowa	Polk	Des Moines	CARPENTER	Carbon monoxide	Parts per million	0.215789	3
5	2018-01-01	Hawaii	Honolulu	Not in a city	Kapolei	Carbon monoxide	Parts per million	0.994737	14
6	2018-01-01	Hawaii	Honolulu	Not in a city	Kapolei	Carbon monoxide	Parts per million	0.2	2
7	2018-01-01	Pennsylvania	Erie	Erie	NaN	Carbon monoxide	Parts per million	0.2	2
8	2018-01-01	Hawaii	Honolulu	Honolulu	Honolulu	Carbon monoxide	Parts per million	0.4	5
9	2018-01-01	Colorado	Larimer	Fort Collins	Fort Collins - CSU - S. Mason	Carbon monoxide	Parts per million	0.3	6

index: The row index, which uniquely identifies each record.
date_local: The date of the record, indicating when the data was collected.
state_name: The name of the state where the data was collected (e.g., "Alaska").
county_name: The name of the county where the data was collected (e.g., "Maniaco").
city_name: The name of the city where the monitoring took place (e.g., "Dulge").
local_site_name: The specific name of the monitoring site (e.g., "BUCKEYE").
parameter_name: The name of the monitored parameter (e.g., "Carbon monoxide"), indicating the type of pollutant being measured.
units_of_measure: The unit of measurement (e.g., "Parts per Million"), indicating how the data is quantified.
arithmetic_mean: The arithmetic mean value of the monitored parameter, representing the average concentration observed over the specified time period.
aqi: The Air Quality Index, which is a composite measure used to indicate overall air quality.

Case study

Sample Questions:

- the 75th percentile for the aqi column?
- the mean value from the aqi column?
- The median value from aqi column?
- The IQR for aqi column?

- What is the 25th percentile for the arithmetic_mean column?
- What is the maximum value from the arithmetic_mean column?
- What is the standard deviation of the arithmetic_mean column?
- What is the mode of the arithmetic_mean column?

- Boxplot of arithmetic_mean by aqi
- Histogram of aqi

Exercise

What Python function can data professionals use to compute the mean, median, and standard deviation all at once?

A. std()

B. median()

C. describe()

D. mean()

Exercise

What two Python functions can you use to compute the range of your dataset?

A. max() and std()

B. max() and median()

C. max() and min()

D. mean() and min()

Content: Introduction to statistics

- Descriptive statistics with Python
- Probability distribution
- Sampling distribution
- Confidence Intervals
- Hypothesis Testing

Probability

- SciPy

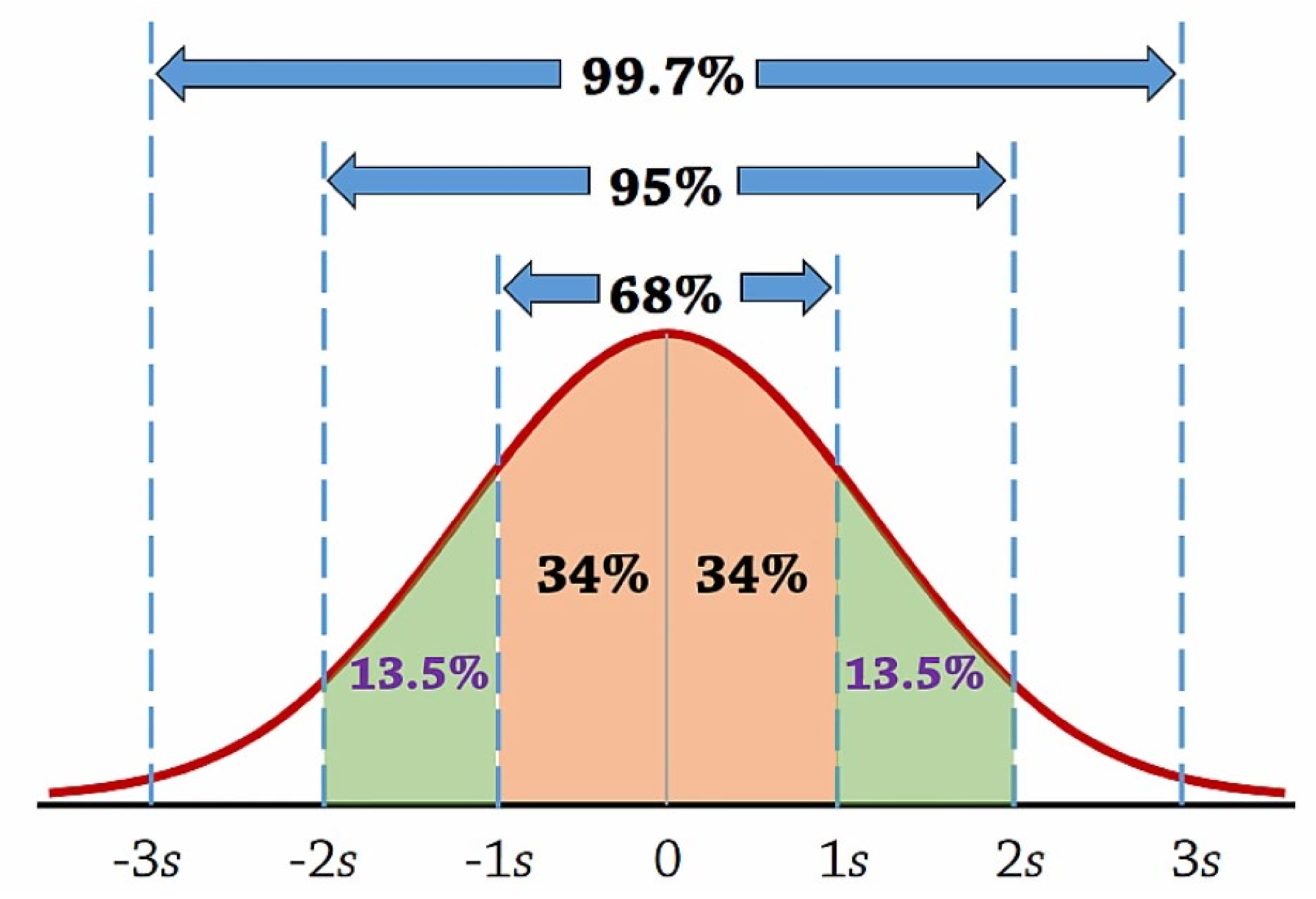
an open-source software for mathematics, science, and engineering. It includes modules for statistics, optimization, integration, linear algebra, Fourier transforms, signal and image processing, ODE solvers, and more.

```
from scipy import stats
```

- .dropna() – remove missing values in the dataset
- .hist() – generate histogram

Probability

- **Normal distribution and Empirical rule**



mean±1SD	68%
mean±2SD	95%
mean±3SD	99.7%

Probability

▶ Normal distribution and Empirical rule

- Calculate Mean and SD

mean±1SD

68%

```
lower_limit = mean_overall_li - 2 * std_overall_li
upper_limit = mean_overall_li + 2 * std_overall_li
((education_districtwise['OVERALL_LI'] >= lower_limit) &
(education_districtwise['OVERALL_LI'] <= upper_limit)).mean()
```

Probability

▶ compute z-score to find outliers

- Calculate the z-score

```
from scipy import stats

education_districtwise['Z_SCORE'] =
stats.zscore(education_districtwise['OVERALL_LI'])
```

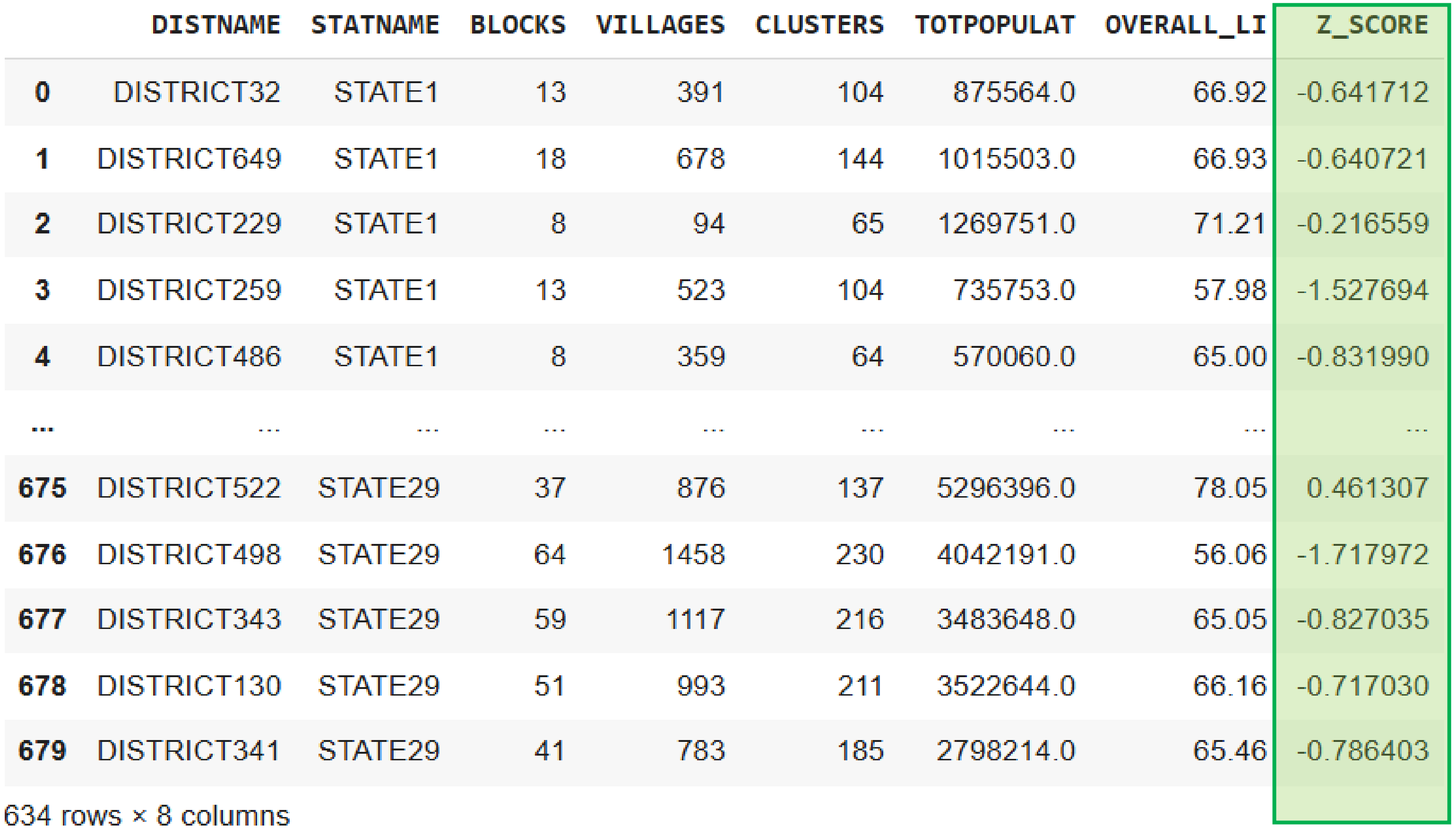
- Find the |z-score|>3

```
education_districtwise[(education_districtwise['Z_SCORE'] > 3)
| (education_districtwise['Z_SCORE'] < -3)]
```

Probability

- **Compute Z-scores to find outliers**

scipy.stats.zscore()



	DISTNAME	STATNAME	BLOCKS	VILLAGES	CLUSTERS	TOTPOPULAT	OVERALL_LI	Z_SCORE
0	DISTRICT32	STATE1	13	391	104	875564.0	66.92	-0.641712
1	DISTRICT649	STATE1	18	678	144	1015503.0	66.93	-0.640721
2	DISTRICT229	STATE1	8	94	65	1269751.0	71.21	-0.216559
3	DISTRICT259	STATE1	13	523	104	735753.0	57.98	-1.527694
4	DISTRICT486	STATE1	8	359	64	570060.0	65.00	-0.831990
...	...	...	...	...	...	...	...	...
675	DISTRICT522	STATE29	37	876	137	5296396.0	78.05	0.461307
676	DISTRICT498	STATE29	64	1458	230	4042191.0	56.06	-1.717972
677	DISTRICT343	STATE29	59	1117	216	3483648.0	65.05	-0.827035
678	DISTRICT130	STATE29	51	993	211	3522644.0	66.16	-0.717030
679	DISTRICT341	STATE29	41	783	185	2798214.0	65.46	-0.786403

634 rows × 8 columns

Z-score could be used to identify values that lie **more than 3 standard deviations below or above the mean**. These values may be considered outliers.

Probability Distribution

- **Implement and visualize the Probability distribution using Python with Matplotlib and SciPy**

	Discrete			Continuous		
Name	**Bernoulli**	**Binomial**	**Poisson**	**Uniform**	**Normal Distribution**	**Exponential**
Notation	$X \sim Bern(p)$	$X \sim b(n,p)$	$X \sim \text{Poisson}(\lambda)$	$X \sim \text{U}\ [a,b]$	$X \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$	$X \sim \exp(\lambda)$
$E(X)$	p	np	λ	$\frac{a+b}{2}$	$\boldsymbol{\mu}$	$\frac{1}{\lambda}$
$\text{Var}(X)$	$p(1-p)$	$np(1-p)$	λ	$\frac{(b-a)^2}{12}$	$\boldsymbol{\sigma}^2$	$\frac{1}{\lambda^2}$

Probability Distribution

- `y = expon.pdf(x, scale=1/lambda_)`
- `y = norm.pdf(x, mu, sigma)`
- `y = uniform.pdf(x, a, b - a)`
- `y = poisson.pmf(x, lambda_)`
- `y = binom.pmf(x, n, p)`

Poisson Distribution

Binomial Distribution

Normal Distribution

Exponential Distribution

Case study 2

`modified_c4_epa_air_quality.csv`

Given that carbon monoxide is a major air pollutant, you will investigate data from the Air Quality Index (AQI) with respect to carbon monoxide.

In this case, one of your main goals is to determine which regions need support to make air quality improvements.

Case study 2

- The aqi_log column represents AQI readings that were transformed logarithmically to suit the objectives of this lab.

Questions:

- How to display number of rows and columns?
- Find out aqi_log fits a specific type of probability distribution
  - Shape: a slight right skew but still appears to be a bell shape
- Use the empirical rule to observe the data, then test and verify that it is normally distributed.
- Use z-score to find outliers
  - What is the potential outliers?

Exercise

A data professional is working with a dataset that has a normal distribution. To test out the empirical rule, they want to find out if roughly 68% of the data values fall within 1 standard deviation of the mean. What Python functions will enable them to compute the mean and standard deviation?

A. mean() and std()

B. mn() and stand()

C. mean() and standard()

D. mn() and std()

Exercise

What Python function is used to compute z-scores for data?

A. median.zscore()
B. stats.zscore()
C. normal.zscore()
D. Mean.zscore()

Content: Introduction to statistics

Descriptive statistics with Python

Probability distribution

Sampling distribution

Confidence Intervals

Hypothesis Testing

Sampling

Use Python to simulate random sampling and make a point estimate of a population mean based on your sample data.

sample() function– simulate random sampling

- n: Refers to the desired sample size
- replace: Indicates whether you are sampling with or without replacement (True: with replacement)
- random_state: Refers to the seed of the random number

Note:

- a random seed is a starting point for generating random numbers.
- Can use arbitrary number to fix the random seed, and give the random number generator a starting point.
- Can use the same random seed to generate the same set of numbers.

Sampling

`education_districtwise.csv`

Collect the data on district literacy rates, and that you have limited time to do so. You can only survey 50 randomly chosen districts, instead of the 634 districts included in your original dataset. The goal of your research study is to estimate the mean literacy rate for *all* 634 districts based on your sample of 50 districts.

Central Limit Theorem

the Central Limit Theorem 中心极限定理

For any population distribution with mean μ and standard deviation σ, the sampling distribution of the sample mean $\bar{X}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$, and the approximation improves as **n** increases.

Notes:

- As the sample increases, the sampling distribution assumes the shape of a **bell curve**.
- If take a large enough sample of the population, the sample mean will be roughly equal to the population mean.

Sampling

▶ The central limit theorem

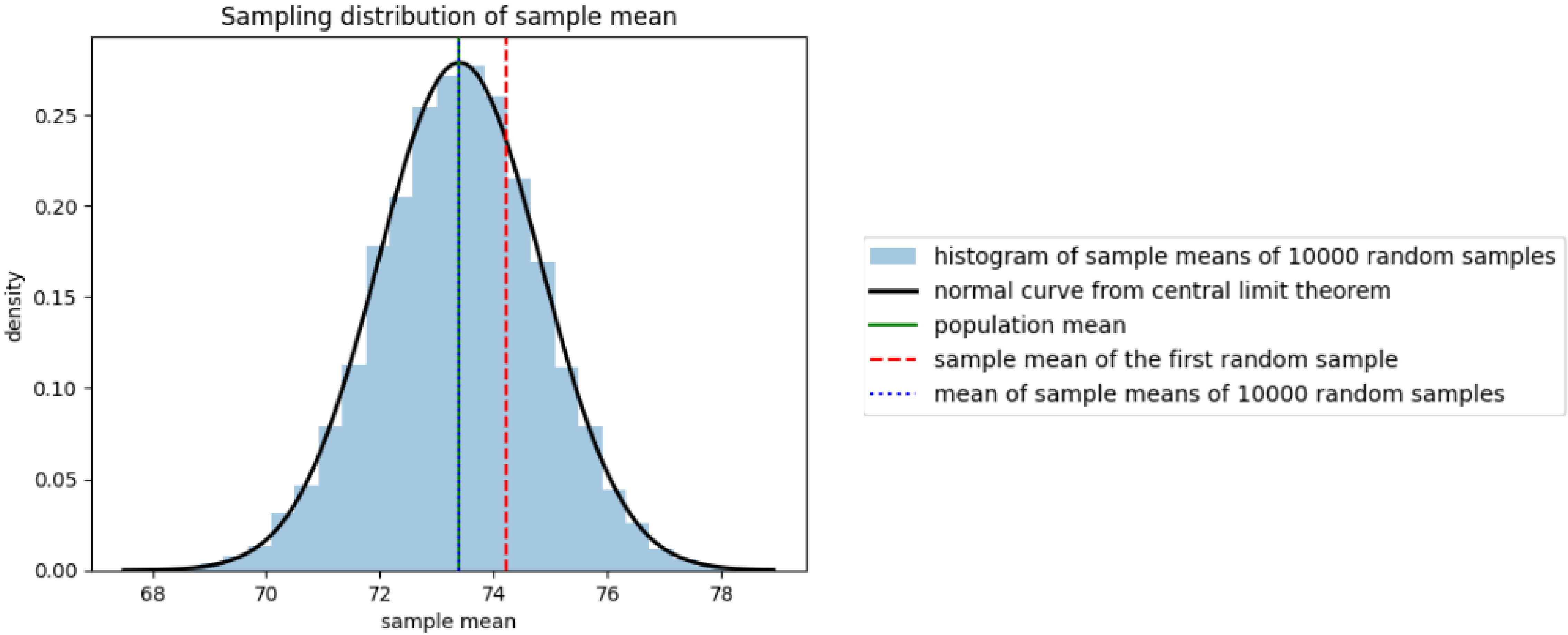
Compute the mean of a sampling distribution with 10,000 samples

1. Create an empty list to store the sample mean from each sample. Name this **estimate_list**.
2. Set up a for-loop with the range()function. The range() function generates a sequence of numbers from 1 to 10,000. The loop will run 10,000 times, and iterate over each number in the sequence.
3. Specify what you want to do in each iteration of the loop. The sample() function tells the computer to take a random sample of 50 districts with replacement–the argument n equals 50, and the argument replace equals True. The append()function adds a single item to an existing list. In this case, it appends the value of the sample mean to each item in the list. Your code generates a list of 10,000 values, each of which is the sample mean from a random sample.
4. Create **a new data frame** for your list of 10,000 estimates. Name a new variable **estimate_df** to store your data frame.

Sampling

- The central limit theorem

Visualize the dataset



Case study

- As a data professional you will often work with extremely large datasets, and utilizing proper sampling techniques helps you improve your efficiency in this work.
- For this activity, you are a member of an analytics team for the Environmental Protection Agency. You are assigned to analyze data on air quality with respect to carbon monoxide—a major air pollutant—and report your findings. The data utilized in this activity includes information from over 200 sites, identified by their state name, county name, city name, and local site name. You will use effective sampling within this dataset.

`c4_epa_air_quality.csv`

Case study

Questions:

- 1. What does the aqi column represent?
- 2. Generate a table of descriptive statistics
  - Based on the preceding table of descriptive statistics, what is the mean value of the aqi column?
  - Based on the preceding table of descriptive statistics, what do you notice about the count value for the aqi column?
- 3. Statistical tests
  - Sample with replacement n = 50, Arbitrary number 42
  - Q1: why the row index 102 repeated twice?
  - Q2: What does random_state do?
  - Q3: Calculate the standard error
- 4. Apply the central limit theorem
  - Repeated times: 10000, sample size = 50
  - Create a new DataFrame
  - Calculate the mean of sampling distribution of 10000 random samples
- 5. Output the distribution using a histogram
- 6. Visualize the relationship between the sampling and normal distributions

Case study

- Answers:
- 1. aqi column represents the Air Quality Index.
- 2. use describe() function
  - mean = 6.757692
  - The count value for the aqi column is 260. This means there are 260 AQI measurements represented in this dataset.
- 3. statistical tests
  - Q1: Sampling with replacement is random
  - Q2: The parameter allows for the reproduction of the same exact sample (i.e., the same set of numbers). This means that the same rows in the dataset will be sampled with replacement each time the command is run.
  - Q3: varies from sample to sample

Case study

What are some key takeaways that you learned from this lab?

- Sampling with replacement on a dataset leads to duplicate rows.
- Sample means are different from population means due to sampling variability.
- The central limit theorem helps describe the sampling distribution of the sample mean for many different types of datasets.

Exercise-sampling

- Which Python function can be used to simulate random sampling?

- A. pandas.DataFrame.mean()
- B. pandas.DataFrame.hist()
- C. pandas.DataFrame.describe()
- D. pandas.DataFrame.sample()

Exercise-sampling

- Which of the following statements describe a **random seed** when specifying random_state in pandas.DataFrame.sample()? Select all that apply.

A. A random seed is a starting point for generating random numbers.

B. Only a negative number may be chosen to fix the random seed.

C. The same random seed may be used over again to generate the same set of numbers.

D. Any non-negative integer can be chosen to fix the random seed.

Content: Introduction to statistics

- Descriptive statistics with Python
- Probability distribution
- Sampling distribution
- Confidence Intervals
- Hypothesis Testing

Confidence Intervals

- Use Python to construct a confidence interval for a point estimate.

Imagine that the department asks you to construct a 95% confidence interval to estimate the mean district literacy rate. You can use Python to construct the confidence interval.

1. Identify a sample statistic
2. Choose a confidence level
3. Find the margin of error
4. Calculate the interval

Confidence Intervals

Construct a 95% confidence interval

```
sampled_data = education_districtwise.sample (n=50,
replace=True, random_state=31208)
```

```
scipy.stats.norm.interval()
```

Confidence Intervals

```
scipy.stats.norm.interval(confidence, loc, scale)
```

confidence:

the confidence level

loc:

the sample mean

scale:

the sample standard error

- 95%
- Mean literacy rate of sample 50 districts
- Estimated standard error

Construct Confidence Interval

- **Construct CI for a mean/proportion**

Confidence level gets higher, the confidence interval gets wider, it is more likely to include the actual population parameter.

As the sample size increases, the confidence interval tends to become narrower, meaning the estimate becomes more precise and the confidence level increases.

Exercise: Confidence Intervals

The Air Quality Index (AQI) is the Environmental Protection Agency's index for reporting air quality. A value close to 0 signals little to no public health concern, while higher values are associated with increased risk to public health. The United States is considering a new federal policy that would create a subsidy for renewable energy in states observing an average AQI of 10 or above.

You've just started your new role as a data analyst in the Strategy division of Ripple Renewable Energy (RRE). RRE operates in the following U.S. states:

California, Florida, Michigan, Ohio, Pennsylvania, Texas.

You've been tasked with constructing an analysis which identifies **which of these states are most likely to be affected**, should the new federal policy be enacted.

Your manager has requested that you do the following for your analysis:

- Provide a summary of the mean AQI for the states in which RRE operates.
- Construct a boxplot visualization for AQI of these states using seaborn.
- Evaluate which state(s) may be most affected by this policy, based on the data and your boxplot visualization.
- Construct a 95% confidence interval for the RRE state with the highest mean AQI.

Content: Introduction to statistics

- Descriptive statistics with Python
- Probability distribution
- Sampling distribution
- Confidence Intervals
- Hypothesis Testing

Hypothesis Testing

Imagine that the Department of Education asks you to collect data on mean district literacy rates for two of the nation’s largest states:

STATE21 and STATE28. STATE28 has almost 40 districts, and STATE21 has more than 70.

Due to limited time and resources, you are only able to survey **20** randomly chosen districts in each state. The department asks you to determine **if the difference between the two mean district literacy rates is statistically significant or due to chance.** This will help the department decide how to distribute government funding to improve literacy. If there is a statistically-significant difference, the state with the lower literacy rate may receive more funding.

you want to test whether the mean literacy rate of STATE21 districts is different from a hypothesized value of 75. (one-sample t-Test)

H_0: The mean literacy rate is equal to the hypothesized value ($\mu = 75$)
H_a: The mean literacy rate is not equal to the hypothesized value ($\mu \neq 75$)

Hypothesis Testing

```
statistic, pvalue =
scipy.stats.ttest_1samp(a=sample data, popmean)
```

- a: The sample observations
- popmean : The expected population mean

Return:

- statistic: The t-statistic value
- pvalue: The p-value for the test

Hypothesis Testing

Imagine that the Department of Education asks you to collect data on mean district literacy rates for two of the nation’s largest states:

STATE21 and STATE28. STATE28 has almost 40 districts, and STATE21 has more than 70.

Due to limited time and resources, you are only able to survey **20** randomly chosen districts in each state. The department asks you to determine **if the difference between the two mean district literacy rates is statistically significant or due to chance.** This will help the department decide how to distribute government funding to improve literacy. If there is a statistically-significant difference, the state with the lower literacy rate may receive more funding.

Use Python to simulate taking a random sample of 20 districts in each state and conduct a **two-sample t-test** based on the sample data.

`Document: education_districtwise.csv`

Hypothesis Testing

```
statistic, pvalue = scipy.stats.ttest_ind(a, b, alternative
equal_var)
```

- a: observations from the first sample
- b: observations from the second sample
- Alternative: defines the alternative hypothesis:
  - ‘two-sided’: mu ≠ popmean
  - ‘less’: mu< popmean
  - ‘greater’: mu>popmean
- equal_var: a Boolean, or true/false statement, which indicates whether the population variance of the two samples is assumed to be equal. To avoid making a wrong assumption, set this argument to False.

Hypothesis Testing

- Select the data of `STATE21` and `STATE28`

```
state21 = education_districtwise[education_districtwise['STATNAME'] == "STATE21"]
state28 = education_districtwise[education_districtwise['STATNAME'] == "STATE28"]
```

- Simulate random sampling

```
sampled_state21 = state21.sample(n=20, replace = True, random_state=13490)
sampled_state28 = state28.sample(n=20, replace = True, random_state=39103)
```

- State H_0 and H_A

H_0:There is no difference in the mean district literacy rates between STATE21 and STATE28.

H_A : There is a difference in the mean district literacy rates between STATE21 and STATE28.

```
tstat, pvalue = stats.ttest_ind(a=sampled_state21['OVERALL_LI'],
b=sampled_state28['OVERALL_LI'], equal_var=False)
```

Case study

You work for an environmental think tank called Repair Our Air (ROA). ROA is formulating policy recommendations to improve the air quality in America, using the Environmental Protection Agency's Air Quality Index (AQI) to guide their decision making. An AQI value close to 0 signals "little to no" public health concern, while higher values are associated with increased risk to public health.

They've tasked you with leveraging AQI data to help them prioritize their strategy for improving air quality in America. ROA is considering the following decisions. For each, construct a hypothesis test and an accompanying visualization, using your results of that test to make a recommendation:

- ROA is considering a metropolitan-focused approach. Within California, they want to know **if the mean AQI in Los Angeles County is statistically different from the rest of California.**
- A new policy will affect those states with **a mean AQI of 10 or greater**. Would Michigan be affected by this new policy?
- With limited resources, ROA has to choose between New York and Ohio for their next regional office. Does New York have a **lower** AQI than Ohio?

Notes:

- For your analysis, you'll default to a 5% level of significance.
- Throughout the lab, for two-sample t-tests, use Welch's t-test (i.e., setting the equal_var parameter to False in scipy.stats.ttest_ind()). This will account for the possibly unequal variances between the two groups in the comparison.

Case study- Questions

Hypothesis 1: Within California, they want to know if the mean AQI in Los Angeles County is statistically different from the rest of California.

Hypothesis 2: A new policy will affect those states with a mean AQI of 10 or greater. Will Michigan be affected by this new policy?

Hypothesis 3: Does New York have a lower AQI than Ohio?

Questions:

- What is your p-value for hypothesis 1, and what does this indicate for your null hypothesis? Did your results show that the AQI in Los Angeles County was statistically different from the rest of California?
- What is your p-value for hypothesis 2, and what does this indicate for your null hypothesis? Will Michigan be affected by the new policy impacting states with a mean AQI of 10 or greater?

Case Study-Explanation

Hypothesis 1: Within California, they want to know if the mean AQI in Los Angeles County is statistically different from the rest of California.

H_0:There is no difference in the mean AQI between Los Angeles and the rest of California.

H_A : There is a difference in the mean AQI between Los Angeles and the rest of California.

```
stats.ttest_ind(a=ca_la['aqi'], b=ca_other['aqi'],
equal_var=False)
```

With a p-value (0.049) being less than 0.05 (as your significance level is 5%), reject the null hypothesis in favor of the alternative hypothesis.Therefore, a metropolitan strategy may make sense in this case.

Case Study-Explanation

Hypothesis 2: A new policy will affect those states with a mean AQI of 10 or greater. Will Michigan be affected by this new policy?

H_0 :The mean AQI of Michigan is less than or equal to 10.

H_A :The mean AQI of Michigan is greater than 10.

```
tstat, pvalue = stats.ttest_1samp(michigan['aqi'],
10, alternative='greater’)
```

With a p-value (0.940) being greater than 0.05 (as your significance level is 5%) and a t-statistic < 0 (-1.74), **fail to reject the null hypothesis**. Therefore, you cannot conclude at the 5% significance level that Michigan's mean AQI is greater than 10. This implies that Michigan would most likely not be affected by the new policy.

Case Study-Explanation

Hypothesis3: With limited resources, ROA has to choose between New York and Ohio for their next regional office. Does New York have a **lower** AQI than Ohio?

H_0: The mean AQI of New York is greater than or equal to that of Ohio.
H_A : The mean AQI of New York is lower than that of Ohio.

```
tstat, pvalue = stats.ttest_ind(a=ny['aqi'],b=ohio['aqi'],
alternative = 'less', equal_var=False)
```

With a p-value (0.03) of less than 0.05, and t-statistic<0 (-2.036), reject the null hypothesis.
Can conclude that 5% significance level that New York has a lower mean AQI than Ohio.