Chapter 1

Introduction to Statistics

We will look at some concepts of Statistics that are fundamental to machine learning

Types of data (numerical, categorical, ordinal)

	Year	Country	Spending_USD	Life_Expectancy
0	1970	AUT	173.391	70.0
1	1970	BEL	149.573	71.0
2	1970	CHE	326.524	73.1
3	1970	DEU	252.311	70.6
4	1970	ESP	82.955	72.0

Year = Numeric/Ordinal
Country = Categorical
Spending_USD = Numeric
Life Expectancy = Numeric

Measures of Central Tendency

Staistic measurement that could help understand the location of data. Forexample average salary.

The famous trio: Mean Median Mode All three provide similar information, however, each one has its merits and demerits. Consider salaries of 100 employees in a start up including CEO. CFO all they way up to analysts. Each of these three could convey a different information. Lets explore these three on our sample data

	mean	median	mode
Spending_USD	2036.384708	1555.264	?

Looking at the results from above, we see a bit of difference between mean and median. Why?

Please try to find out the mode for this variable as an exercise.

Measures of Dispersion

Variance

- Variance is the average of the squared differences from the mean of a set of data.

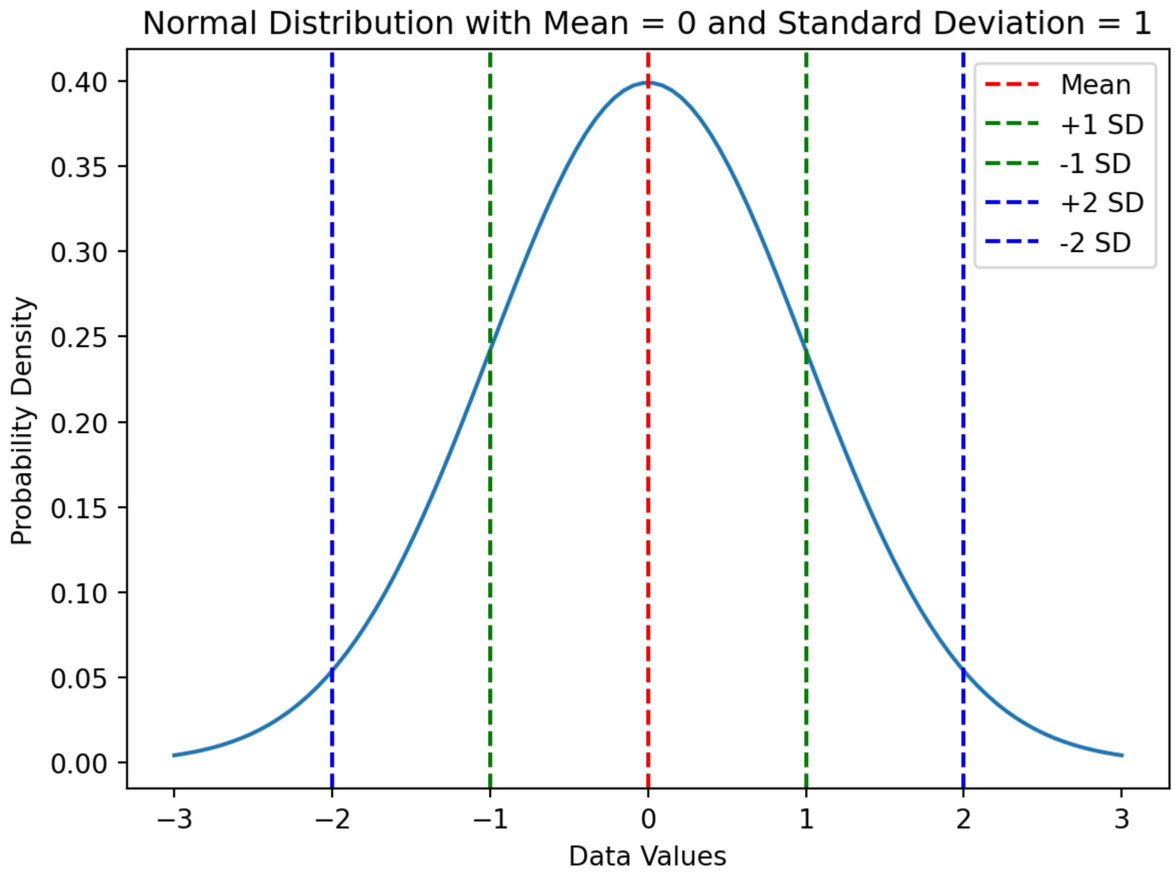
- It tells you how much, on average, each data point deviates from the mean

Standard deviation

- Standard deviation is square root of variance
- It tells you how much, on average, each data point deviates from the mean, but in the same units as the original data

Both measures tell you how spread out a set of data is, but standard deviation is generally easier to interpret and compare across different datasets

Lets see this through a normal curve



Range

Range is nothing but spread between the highest and lowest values i.e., min - max

Lets see how to find these measures quickly using pandas

```
# import pandas as pd

mydata["Spending_USD"].describe()
```

```
count     1556.000000
mean      2036.384708
std       1670.910567
min         18.093000
25%        781.457000
50%       1555.264000
75%       2876.931500
max      11859.179000
Name: Spending_USD, dtype: float64
```

Sampling

Population: In statistics, a set of observations under study is considered as a population

Sample: Examining the entire population is impractical and impossible due to size, cost etc. So we select a subset of observations from the population(Universe) to ascertain estimates for key statistics like mean.

Types of Sampling: Probability Sampling: Every member of the population has a known chance of being selected, ensuring representativeness.
Examples include simple random sampling, stratified sampling

Non-probability Sampling: Selection is not based on random chance, making it potentially less representative but sometimes more convenient

Sample Size:

Choosing the right sample size is crucial for accurate inferences. Larger samples are generally more reliable, but also require more resources

Applications: Sampling is used in various fields like marketing research (surveying consumer preferences), public opinion polls, medical studies (testing drug efficacy), quality control procedures, and many more.

Simple Random Sampling

Note that it could be sampling with replacement or without replacement

```python
import random

# Define your population (list of items)
population = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

# Specify the desired sample size
sample_size = 5

# Use the random.sample() function to draw a simple random sample
random_sample = random.sample(population, sample_size)

# Print the selected sample
print("Random sample:", random_sample)
```

```
Random sample: [1, 2, 10, 3, 8]
```

Stratified Random Sampling

Imagine you want to survey 100 students from a school of 400 to understand their opinions on school lunches. The student body comprises 250 males and 150 females. To ensure representation of both genders in your sample, you'd use stratified sampling:

1. Divide the population into strata: Divide the students into two strata based on gender: male and female.

2. Determine sample size for each stratum: Calculate the proportion of each stratum in the population:

   - Males: 250/400 = 0.625 (62.5%)
   - Females: 150/400 = 0.375 (37.5%)
   - Allocate sample sizes proportionally: 62.5% of 100 = 63 males
     37.5% of 100 = 37 females

3. Draw random samples from each stratum: Use simple random sampling within each stratum to select the specified number of students:

   - Randomly select 63 males from the male stratum.
   - Randomly select 37 females from the female stratum.

Probability

Probability is a fascinating concept that deals with the likelihood of an event happening. It's essentially a numerical way of expressing how certain we are about something occurring.

Conditional Probability

Imagine you have a bag containing 10 marbles: - 5 red marbles - 3 blue marbles - 2 green marbles

Probability distributions

- Conditional probability deals with the probability of an event happening given that another event has already occurred

- It involves updating the sample space to reflect the new information.

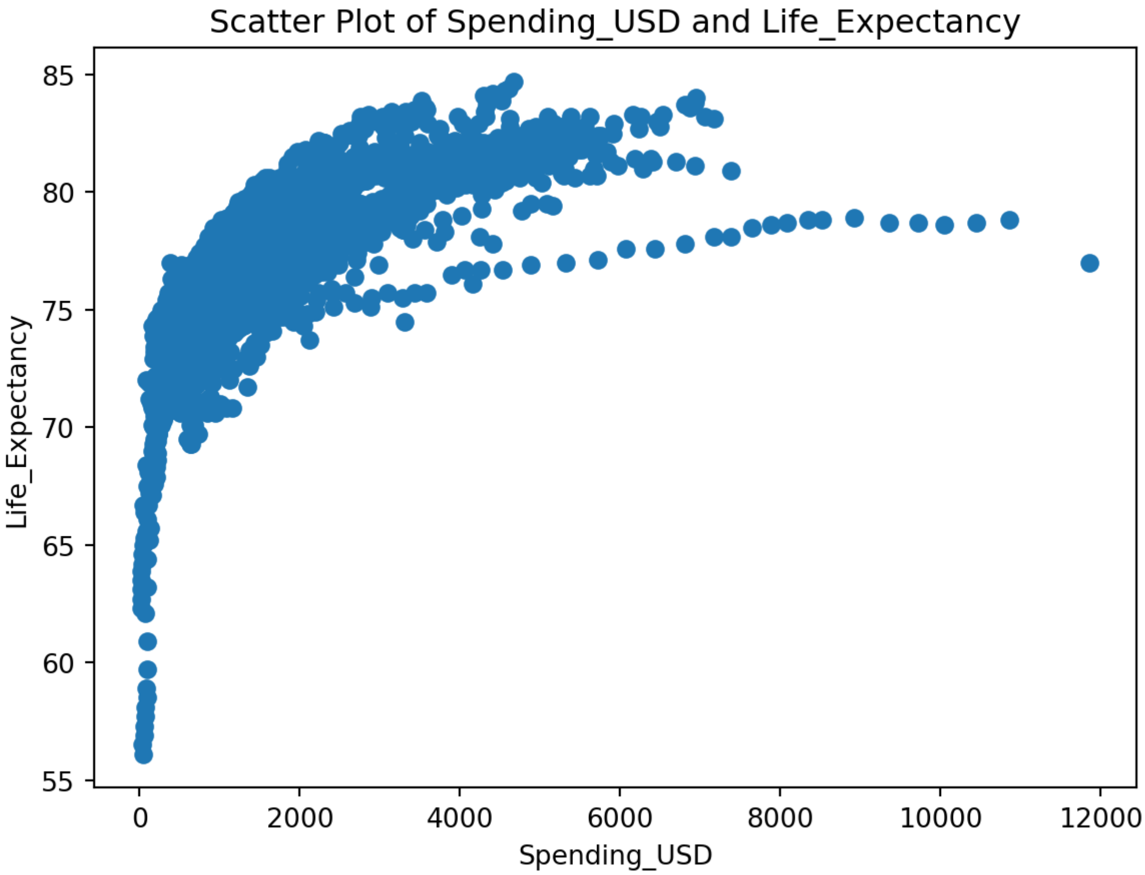
- The formula for conditional probability is:

```
P(A | B) = P(A and B) / P(B)
```

Correlation and Covariance

Correlation refers to a statistical relationship between two variables. It basically tells you how closely the two variables change together.

```
Correlation coefficient between Spending_USD and Life_Expectancy : 0.7160824094352211
```



Covariance, like correlation, measures the relationship between two random variables. However, while correlation gives you the direction and strength of the relationship (think of it as the compass), covariance tells you the magnitude and direction (think of it as the distance and direction on the map).

Think of temperature and ice cream sales - as temperature goes up, ice cream sales tend to go down

Hypothesis Testing

Hypothesis testing is a fundamental statistical technique used to draw conclusions about a population based on limited data from a sample. It involves formulating a hypothesis (a claim about a population parameter) and then using statistical methods to assess whether the data provides enough evidence to reject or support that hypothesis.

1. Formulate the hypothesis:
   - Null hypothesis (H0): This is the default assumption, typically stating no difference or no effect between variables.
   - Alternative hypothesis (Ha): This is the opposite of the null hypothesis, representing the desired effect or difference you want to observe.
2. Collect data:
   - Obtain a representative sample from the population of interest.
3. Choose a statistical test:
   - Select a test appropriate for the type of data, hypothesis, and desired outcome.
4. Conduct the test:

   - Analyze the sample data using the chosen statistical test. This typically involves calculating a test statistic and comparing it to a pre-defined critical value.
5. Interpret the results:
   - Based on the test statistic and critical value, you can:
      - Reject the null hypothesis: If the data provides strong evidence against H0, you can conclude that Ha is likely true.
      - Fail to reject the null hypothesis: This doesn’t necessarily mean H0 is true, but simply that the evidence was not strong enough to reject it.

Important considerations:

Significance level (alpha): This sets the probability of falsely rejecting the null hypothesis (type I error). Typically set at 0.05 (5%).

Statistical power: The probability of correctly rejecting the null hypothesis when it’s actually false. Ideally, you want high power to avoid missing an actual effect.

Assumptions: Many statistical tests rely on certain assumptions about the data (e.g., normality, independence). Ensure your data meets these assumptions for valid results.

Applications of hypothesis testing:

Scientific research: Testing the effectiveness of new drugs, comparing different teaching methods, analyzing the impact of interventions.

Business analysis: Evaluating marketing campaigns, comparing product features, assessing quality control processes.

Social science research: Investigating relationships between variables, testing theoretical models, analyzing demographic trends.

P-Values and Significance Levels

Suppose you are flipping a coin. The null hypothesis (H0) is that the coin is fair, meaning that there is a 50% chance of heads and a 50% chance of tails. The alternative hypothesis (Ha) is that the coin is not fair, meaning that the probability of heads is not equal to 50%.

You flip the coin 10 times and get 7 heads. The p-value is the probability of getting 7 or more heads if the null hypothesis is true. In this case, the p-value is 0.064.

The significance level (α) is the probability of rejecting the null hypothesis when it is actually true. The most common significance level is 0.05.

If the p-value is less than the significance level, then the result is considered statistically significant. In this case, the p-value (0.064) is greater than the significance level (0.05), so the result is not statistically significant.

In other words, there is a 6.4% chance of getting 7 or more heads if the coin is actually fair. Since this probability is greater than the significance level, we cannot reject the null hypothesis. We do not have enough evidence to conclude that the coin is not fair.

However, if the p-value had been less than the significance level (e.g., 0.02), then we would have rejected the null hypothesis. We would have concluded that there is enough evidence to suggest that the coin is not fair.

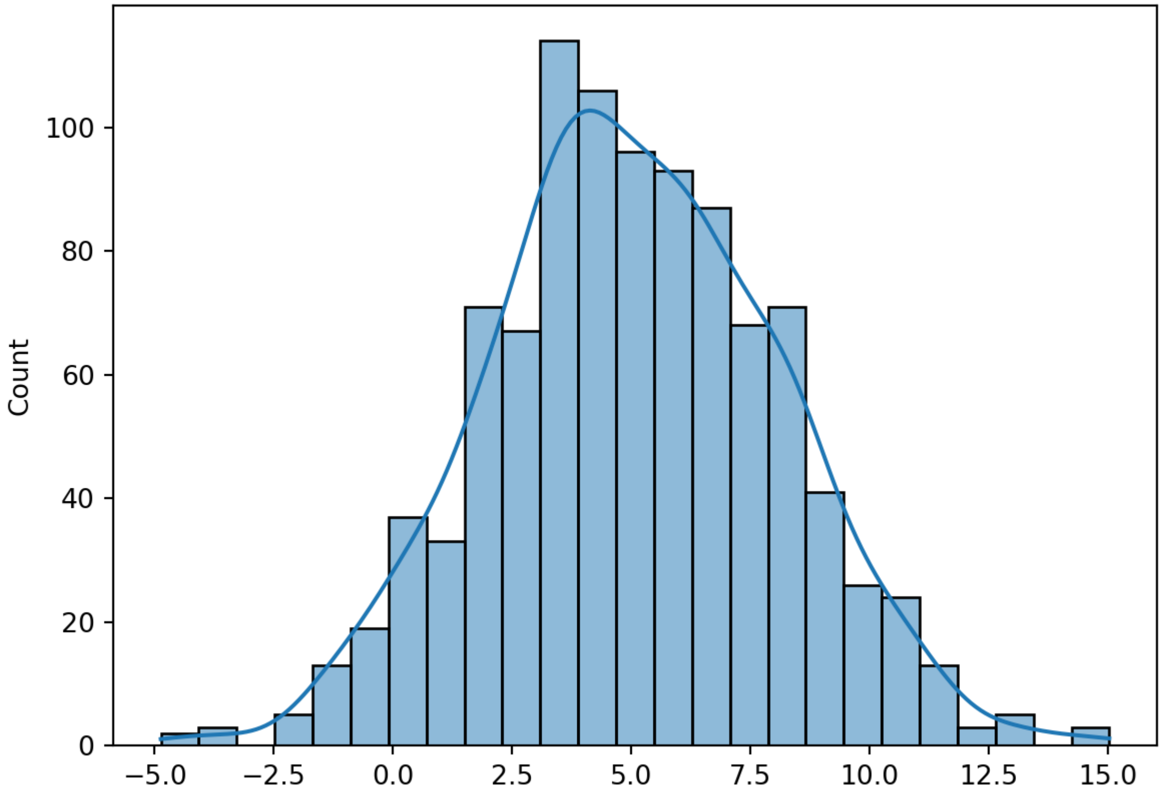
In summary, the p-value is the probability of getting a result as extreme as or more extreme than the one you observed, assuming that the null hypothesis is true. The significance level is the probability of rejecting the null hypothesis when it is actually true. If the p-value is less than the significance level, then the result is considered statistically significant.

```
# Importing packages
import numpy as np
import seaborn as sns

# creating a sample of 10000 random values with mean 5 and standard deviation 3
value = np.random.normal(loc=5,scale=3,size=1000)

# Plot the histogram and normal curve
sns.histplot(value, kde = True)
```

```
<Axes: ylabel='Count'>
```


Count