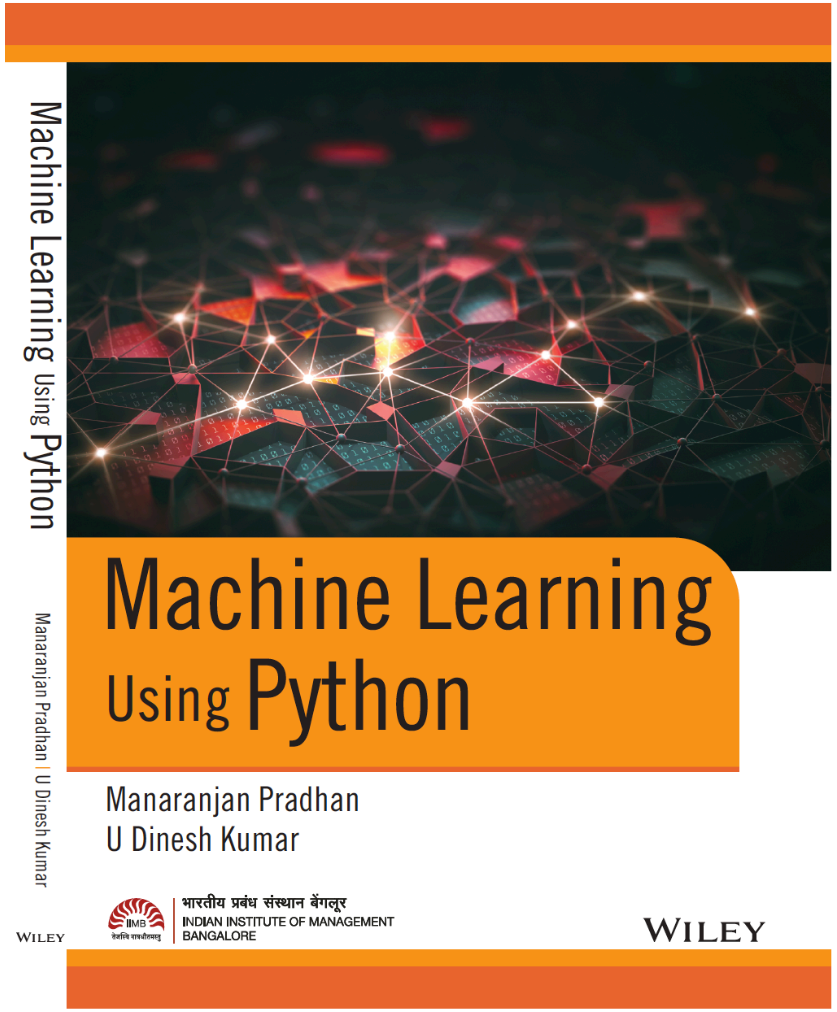

Machine Learning Using Python
Manaranjan Pradhan
U Dinesh Kumar
भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE
WILEY

Chapter 03: Probability Distributions and Hypothesis Tests

Prepared By: Purvi Tiwari

Learning Objectives

- Learn the concepts of a random variable and its role in analytics.
- Understanding different probability distributions and their applications.
- Deriving insights from statistical measures such as mean, variable, probability distribution functions, confidence interval etc.
- Learning hypothesis tests such as one-sample Z-test, t-test, two-sample t-test, paired t-test, chi-square tests, and analysis of variance (ANOVA).

Probability Theory - Terminology

- **Random Experiment –** an experiment in which the outcome is not known with certainty. That is, the output of a random experiment cannot be predicted with certainty.
- **Sample Space –** the universal set that consists of all possible outcomes of an experiment and individual outcomes are called the elementary events.

✓**Experiment:** Outcome of a college application

Sample Space: S = {admitted, not admitted}

✓**Experiment:** Predicting customer churn at an individual customer level

Sample Space: S = {Churn, No Churn}

Probability Theory – Terminology (Cntd.)

- **Event –** a subset of a sample space and probability is usually calculated with respect to an event. For Example:

1. Number of cancellation of orders placed at an E-commerce portal site exceeding 10%
2. The number of fraudulent credit card transactions exceeding 1%
3. The life of a capital equipment being less than one year.

Random Variables

- A random variable is a function that maps every outcome in the sample space to a real number. It can both be discrete and continuous.
- Discrete random variable – If the random variable X can assume only a finite or countably infinite set of values, then it is called a discrete random variable. Examples:

1. Credit rating (low, medium, and high credit rating)
2. Customer churn (churn and do not churn)
3. Fraud (fraudulent transaction and genuine transaction)

- Discrete random variables are described using probability mass function (PMF) and cumulative distribution function (PDF).

Random Variables (Cntd.)

- Continuous random variable – A random variable X which can take a value from an infinite set of values is called a continuous random variable. Examples:

1. Market share of a company (any value between 0 and 100%).
2. Percentage of attrition of employees of an organization.
3. Time-to-failure of an engineering system.

- Continuous random variables are described using probability density function (PDF) and cumulative distribution function (CDF).

Binomial Distribution

- A discrete probability distribution function
- A random variable X is said to follow a binomial distribution when:

1. Random variable can have only two outcomes – success and failure
2. Objective is to find the probability of getting x successes out of n trials
3. Probability of success is p and probability of failure is (1-p)
4. Probability p is constant and does not change between trials

Binomial Distribution (Cntd.)

- Few examples of business problems:

1. Customer churn where the outcomes are (churn and no churn)
2. Fraudulent insurance claims where the outcomes are (Fraudulent claim and Genuine claim)
3. Loan repayment default by a customer where the outcomes are Default and No default)

Binomial Distribution (Cntd.)

- PMF of the binomial distribution

$$PMF(x) = P(X = x) = \binom{n}{x} \times p^{x} \times (1 - p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- CDF of a binomial distribution

$$CDF(x) = P(X \le x) = \sum_{k=0}^{x} \binom{n}{k} \times p^{k} \times (1 - p)^{n-k}$$

Binomial Distribution (Cntd.) Example:

- Fashion Trends Online (FTO) is an e-commerce company that sells women apparel. It is observed that 10% of their customers return the items purchased by them for many reasons (such as size, color, and material). On a specific day, 20 customers purchased items from FTO. Calculate:

1. Probability that exactly 5 customers will return the items
2. Probability that a maximum of 5 customers will return the items
3. Probability that more than 5 customers will return the items
4. Average number of customers who are likely to return the items and the variance and the standard deviation of the number of returns.

Binomial Distribution (Cntd.)

Solution:

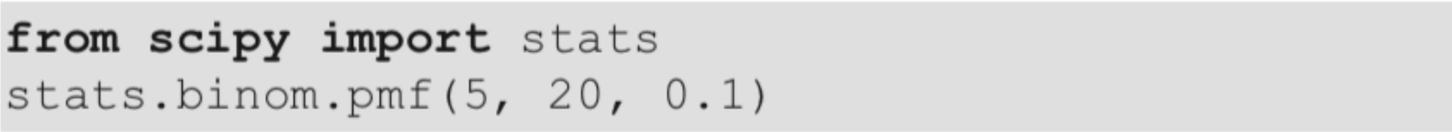
1. Probability that exactly 5 customers will return the items

Solution:

(a) Expected number of successful trials (5)

(b) Total number of trials (20)

(c) The probability of success (0.1)



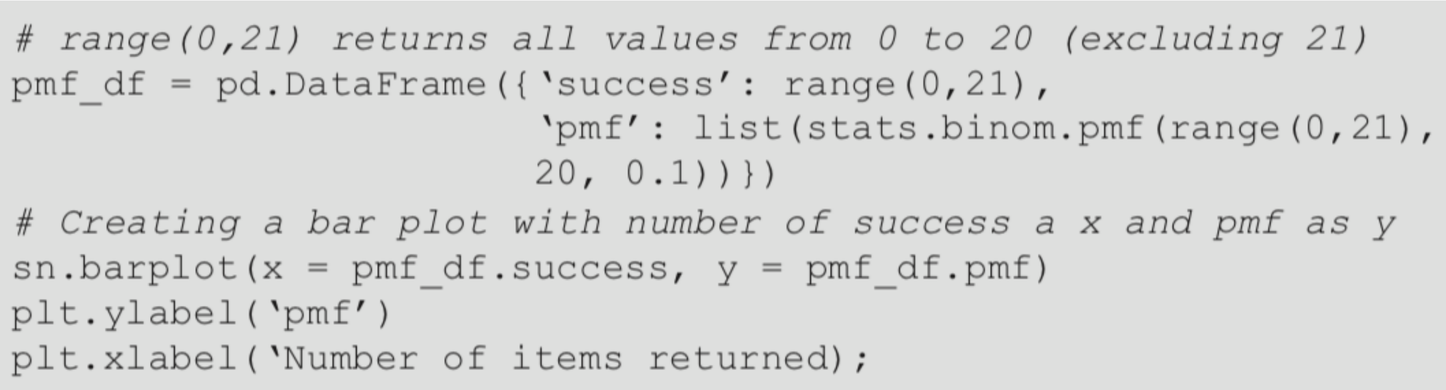
```
from scipy import stats
stats.binom.pmf(5, 20, 0.1)
```

The corresponding probability is 0.03192, that is, the probability that exactly 5 customers will return the items is approximately 3%.

Binomial Distribution (Cntd.)

Solution:

- To visualize how the PMF varies with number of successful trials



```
# range(0,21) returns all values from 0 to 20 (excluding 21)
pmf_df = pd.DataFrame({'success': range(0,21),
                       'pmf': list(stats.binom.pmf(range(0,21),
                       20, 0.1))})
# Creating a bar plot with number of success a x and pmf as y
sn.barplot(x = pmf_df.success, y = pmf_df.pmf)
plt.ylabel('pmf')
plt.xlabel('Number of items returned);
```

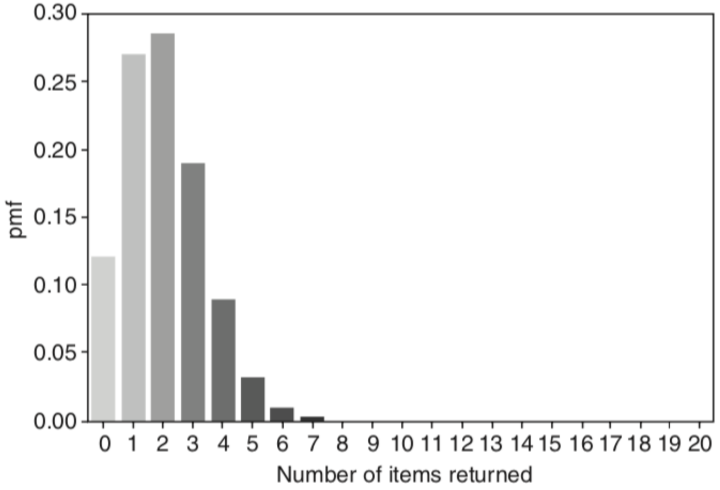


FIGURE 3.1 Binomial distribution.

Binomial Distribution (Cntd.)

Solution:

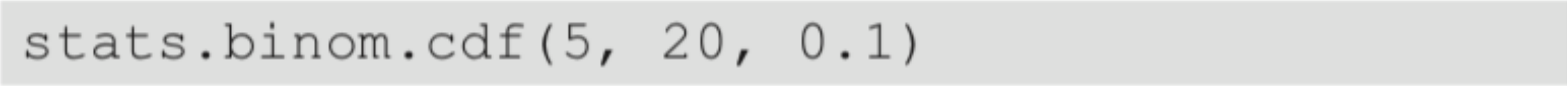
2. Probability that a maximum of 5 customers will return the items

Solution:

(a) Expected number of successful trials (5)

(b) Total number of trials (20)

(c) The probability of success (0.1)



```
stats.binom.cdf(5, 20, 0.1)
```

The corresponding probability is 0.9887.

Binomial Distribution (Cntd.)

Solution:

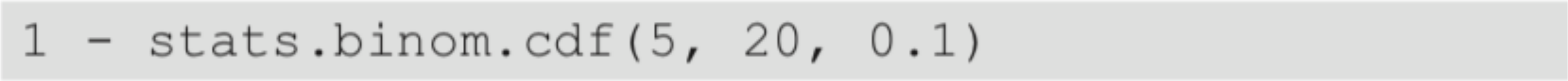
3. Probability that more than 5 customers will return the items

Solution:

(a) Expected number of successful trials (5)

(b) Total number of trials (20)

(c) The probability of success (0.1)



```
1 - stats.binom.cdf(5, 20, 0.1)
```

The corresponding probability is 0.0112.

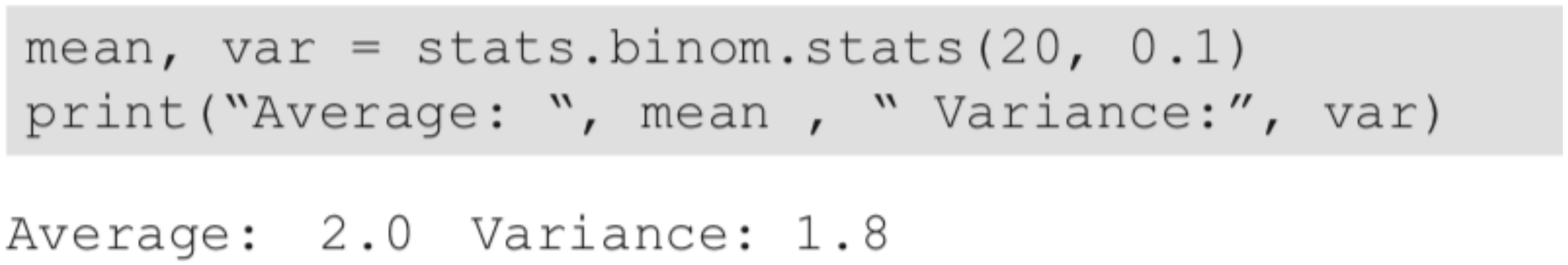
Binomial Distribution (Cntd.)

Solution:

4. Average number of customers who are likely to return the items and the variance and the standard deviation of the number of returns.

(a) Average of a binomial distribution is given by n*p

(b) Variance of the binomial distribution is n*p*(1-p)



```
mean, var = stats.binom.stats(20, 0.1)
print(“Average: “, mean , “ Variance:”, var)
```

```
Average:  2.0  Variance: 1.8
```

Poisson Distribution

- Few examples of business problems:
    1. Number of cancellation of orders by customers at an e-commerce portal
    2. Number of customer complaints
    3. Number of cash withdrawals at an ATM
    4. Number of typographical errors in a book
    5. Number of potholes on the roads

Poisson Distribution (Cntd.)

- The PMF of a Poisson distribution is given by

$$P(X=k)=\frac{e^{-\lambda}\times\lambda^{k}}{k!}$$

Where λ is the rate of occurrence of the events per unit of measurement (in many situations the unit of measurement is likely to be time)

Poisson Distribution (Cntd.)

Example:

- The number of calls arriving at a call center follows a Poisson distribution at 10 calls per hour.
    1. Calculate the probability that the number of calls will be maximum 5.
    2. Calculate the probability that the number of calls over a 3-hour period will exceed 30.

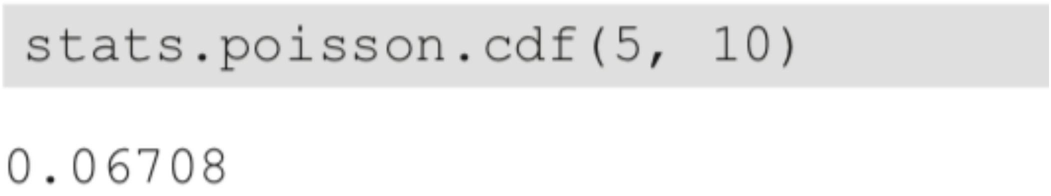
Poisson Distribution (Cntd.)

Solution:

1. Calculate the probability that the number of calls will be maximum 5.

   a) First parameter: Number of events (5 calls) for which the probability needs to be calculated.

   b) Second parameter: The average numbers of events (10 calls per hour)



```
stats.poisson.cdf(5, 10)
```

```
0.06708
```

The corresponding probability is 0.067

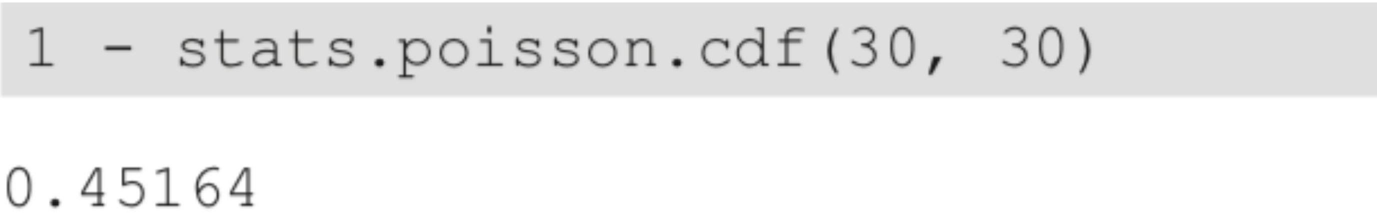
Poisson Distribution (Cntd.)

Solution:

2. Calculate the probability that the number of calls over a 3-period will exceed 30.

   a) Average calls per hour is 10

   b) The mean number of calls over 3 hours is 30



```
1 - stats.poisson.cdf(30, 30)
```

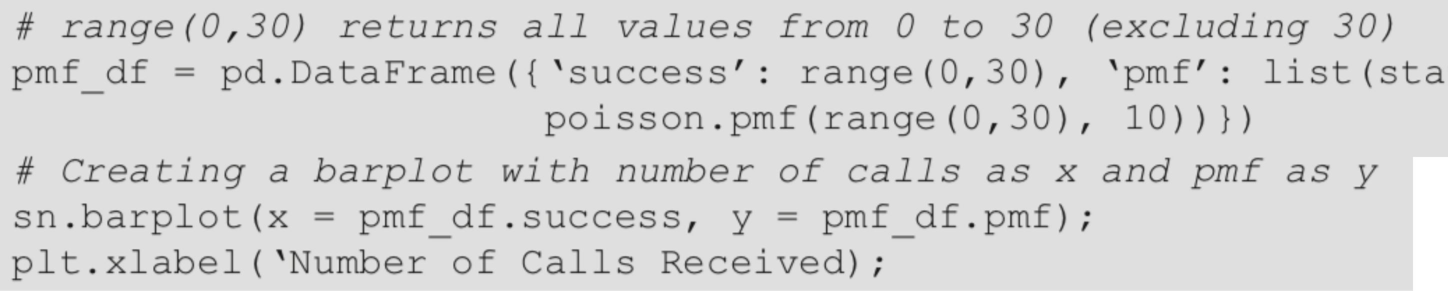
```
0.45164
```

The corresponding probability is 0.451

Poisson Distribution (Cntd.)

Solution:

Plotting PMF for all possible number of calls the call center can receive ranging from 0 to 30.



```
# range(0,30) returns all values from 0 to 30 (excluding 30)
pmf_df = pd.DataFrame({‘success’: range(0,30), ‘pmf’: list(stats.
                       poisson.pmf(range(0,30), 10))})
# Creating a barplot with number of calls as x and pmf as y
sn.barplot(x = pmf_df.success, y = pmf_df.pmf);
plt.xlabel(‘Number of Calls Received);
```

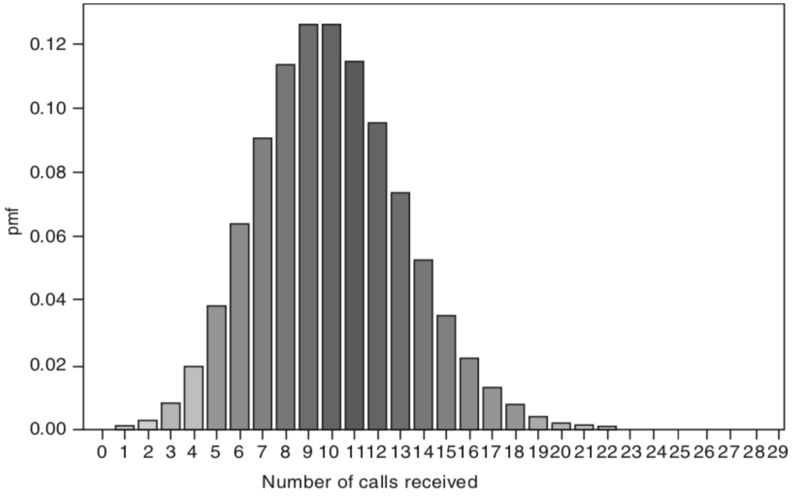


FIGURE 3.2 Poisson distribution.

Exponential Distribution

- A single parameter continuous distribution.
- A process in which events occur continuously and independentally at a constant average rate.
- The probability density function is given by

$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

- The parameter λ is the scale parameter and represents the rate of occurrence of the event
- Mean of exponential distribution is given by $1/\lambda$

Exponential Distribution

Example:

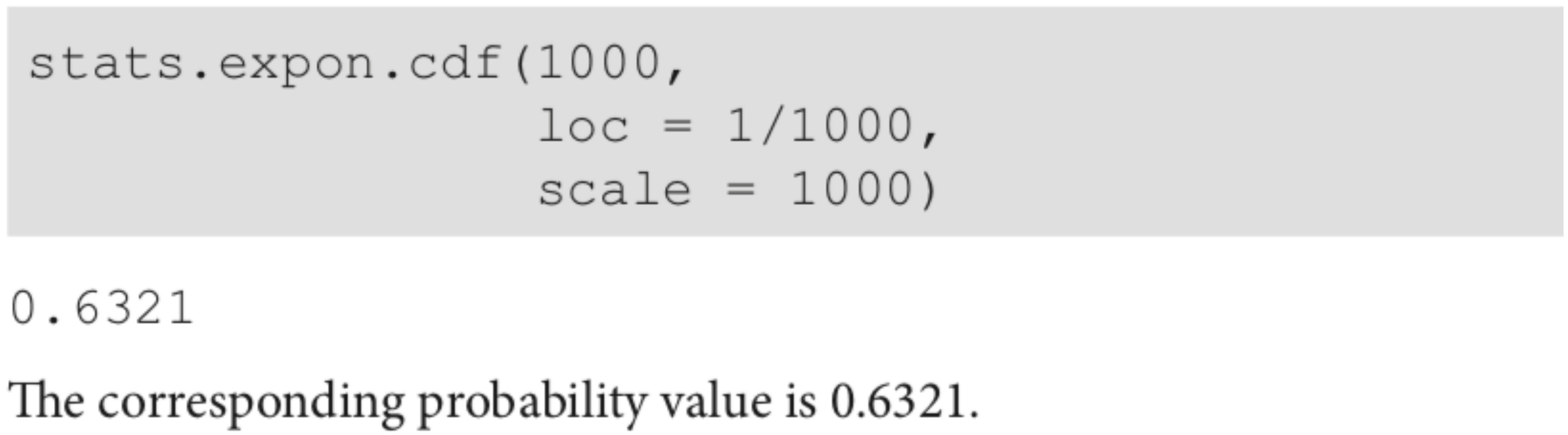
- The time-to-failure of an avionic system follows an exponential distribution with a mean time between failures (MTBF) of 1000 hours. Calculate

1. The probability that the system will fail before 1000 hours.
2. The probability that it will not fail upto 2000 hours.
3. The time by which 10% of the system will fail (i.e., calculate P10 life)

Exponential Distribution (Cntd.)

Solution:

1. The probability that the system will fail before 1000 hours.



```
stats.expon.cdf(1000,
                loc = 1/1000,
                scale = 1000)
```

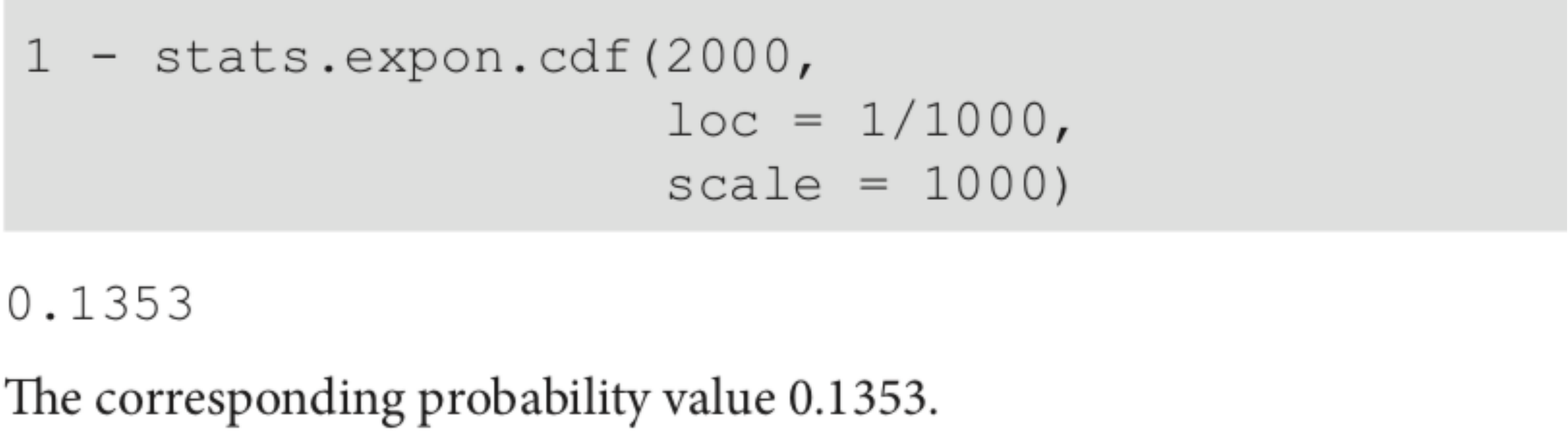
```
0.6321
```

The corresponding probability value is 0.6321.

Exponential Distribution (Cntd.)

Solution:

2. The probability that the system will not fail upto 2000 hours.



```
1 - stats.expon.cdf(2000,
                    loc = 1/1000,
                    scale = 1000)
```

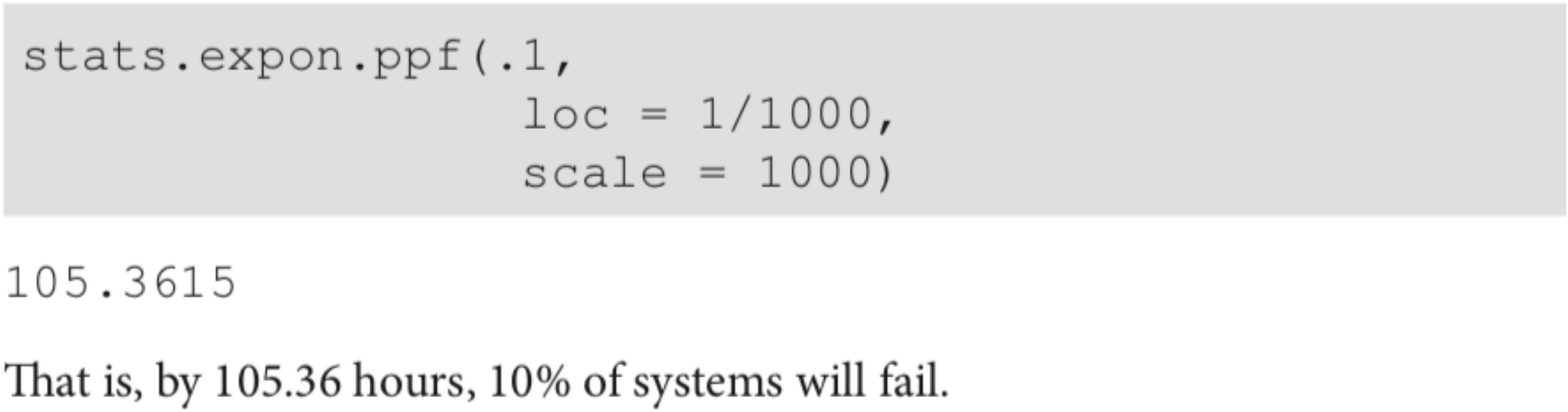
```
0.1353
```

The corresponding probability value 0.1353.

Exponential Distribution (Cntd.)

Solution:

3. The time by which 10% of the system will fail (i.e., P10 life).



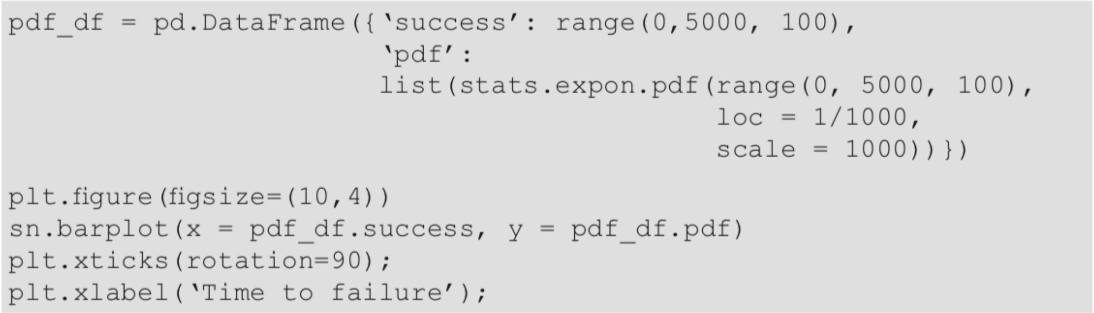
```
stats.expon.ppf(.1,
                loc = 1/1000,
                scale = 1000)
```

```
105.3615
```

That is, by 105.36 hours, 10% of systems will fail.

Exponential Distribution (Cntd.)

Plotting the PDF function against different time-to-failure hours.



```
pdf_df = pd.DataFrame({'success': range(0,5000, 100),
                       'pdf':
                       list(stats.expon.pdf(range(0, 5000, 100),
                                            loc = 1/1000,
                                            scale = 1000))})

plt.figure(figsize=(10,4))
sn.barplot(x = pdf_df.success, y = pdf_df.pdf)
plt.xticks(rotation=90);
plt.xlabel('Time to failure');
```

The corresponding exponential distribution is shown in Figure 3.3.

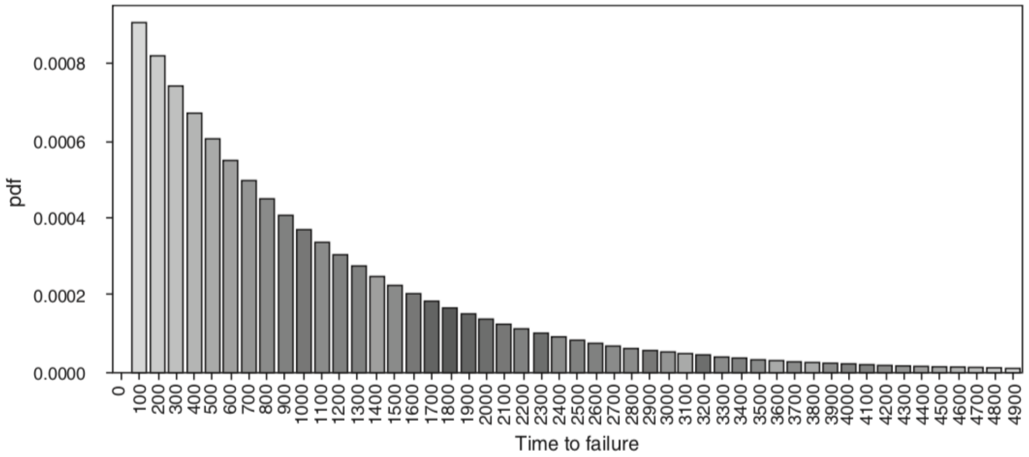


FIGURE 3.3 Exponential distribution.

Normal Distribution

- Also known as Gaussian distribution
- A continuous distribution
- Normal distribution is observed across many naturally occurring measures such as age, salary, sale volume, birth weight, height, etc.
- Popularly known as bell curve

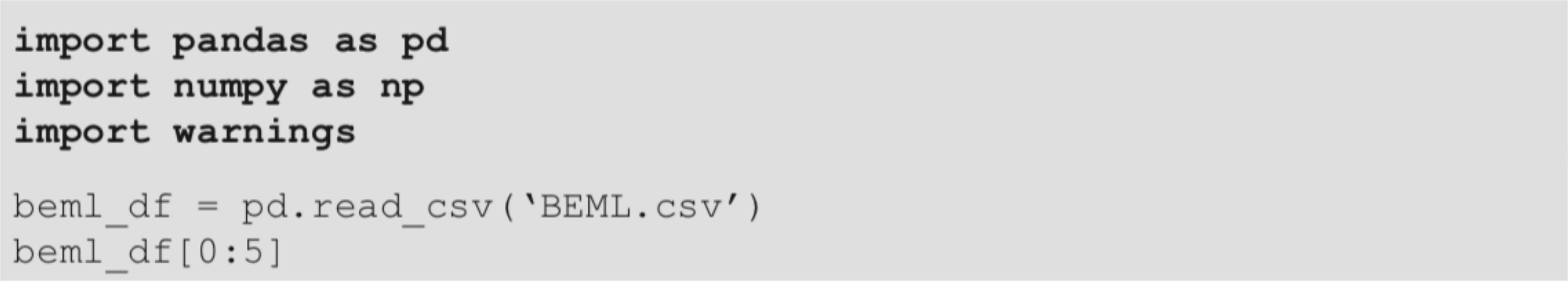
Normal Distribution (Cntd.)

Example:

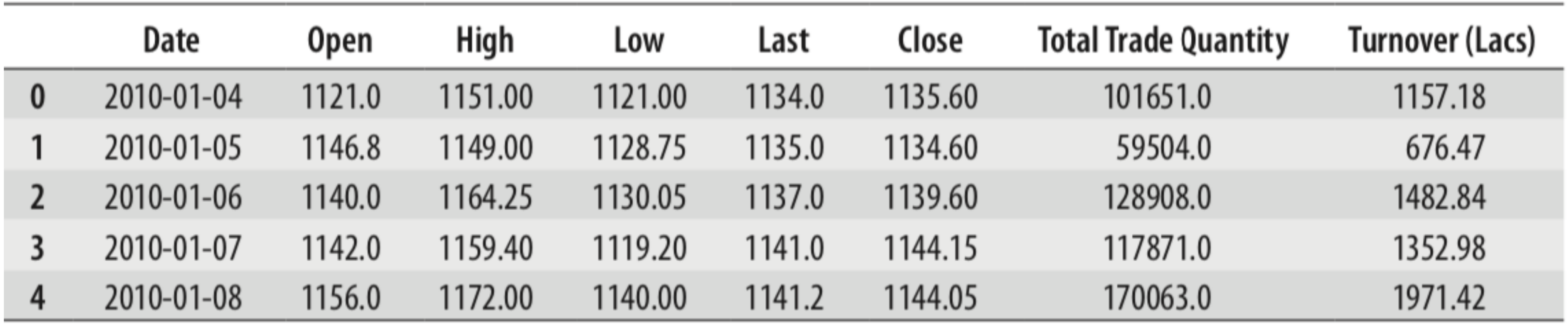
- Imagine a scenario where an investor wants to understand the risks and returns associated with various stocks before investing in them.
- We will evaluate two stocks: BEML and GLAXO.
- The daily trading data for each stock is taken for the period starting from 2010 to 2016 from BSE site.
- Reference: (www.bseindia.com)

Normal Distribution (Cntd.)

Solution: loading the data



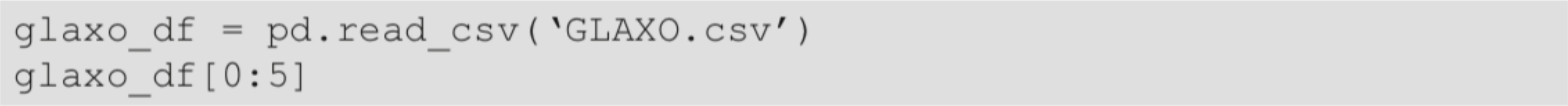
```
import pandas as pd
import numpy as np
import warnings

beml_df = pd.read_csv(‘BEML.csv’)
beml_df[0:5]
```



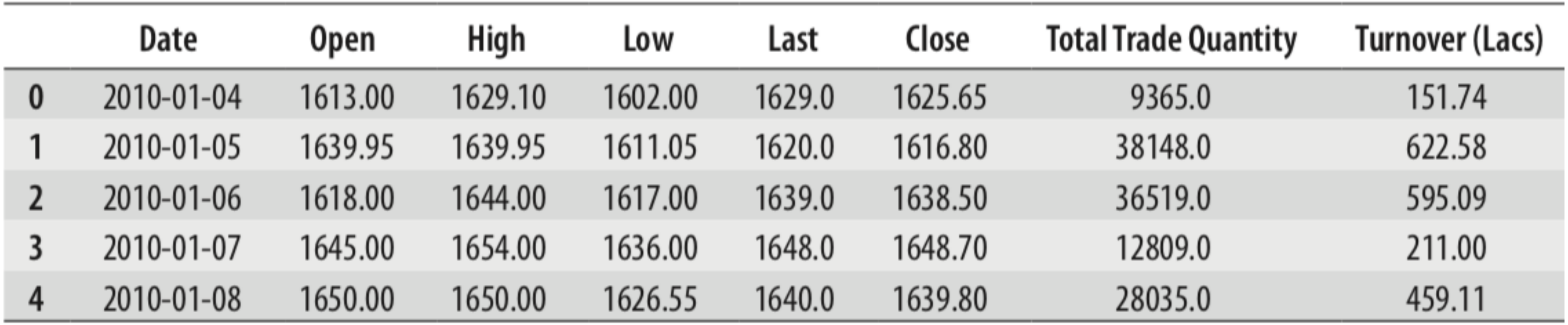
	Date	Open	High	Low	Last	Close	Total Trade Quantity	Turnover (Lacs)
0	2010-01-04	1121.0	1151.00	1121.00	1134.0	1135.60	101651.0	1157.18
1	2010-01-05	1146.8	1149.00	1128.75	1135.0	1134.60	59504.0	676.47
2	2010-01-06	1140.0	1164.25	1130.05	1137.0	1139.60	128908.0	1482.84
3	2010-01-07	1142.0	1159.40	1119.20	1141.0	1144.15	117871.0	1352.98
4	2010-01-08	1156.0	1172.00	1140.00	1141.2	1144.05	170063.0	1971.42

Normal Distribution (Cntd.)

Solution: loading the data



```
glaxo_df = pd.read_csv('GLAXO.csv')
glaxo_df[0:5]
```

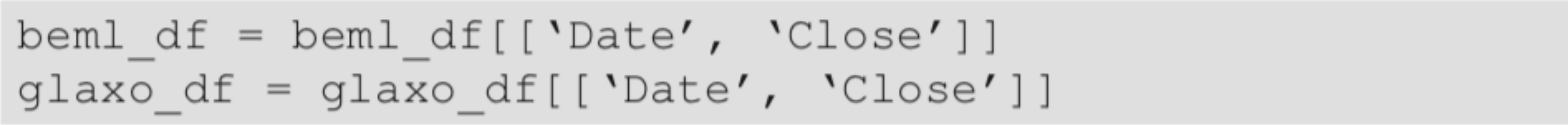


	Date	Open	High	Low	Last	Close	Total Trade Quantity	Turnover (Lacs)
0	2010-01-04	1613.00	1629.10	1602.00	1629.0	1625.65	9365.0	151.74
1	2010-01-05	1639.95	1639.95	1611.05	1620.0	1616.80	38148.0	622.58
2	2010-01-06	1618.00	1644.00	1617.00	1639.0	1638.50	36519.0	595.09
3	2010-01-07	1645.00	1654.00	1636.00	1648.0	1648.70	12809.0	211.00
4	2010-01-08	1650.00	1650.00	1626.55	1640.0	1639.80	28035.0	459.11

Normal Distribution (Cntd.)

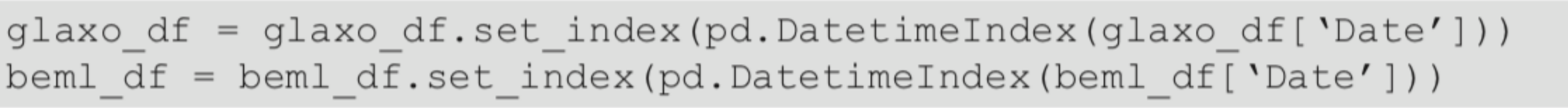
Solution:

- Selecting Date and Close columns from the DataFrames, since the analysis will involve only daily prices.



```
beml_df = beml_df[[‘Date’, ‘Close’]]
glaxo_df = glaxo_df[[‘Date’, ‘Close’]]
```

- Setting the Datetime Index

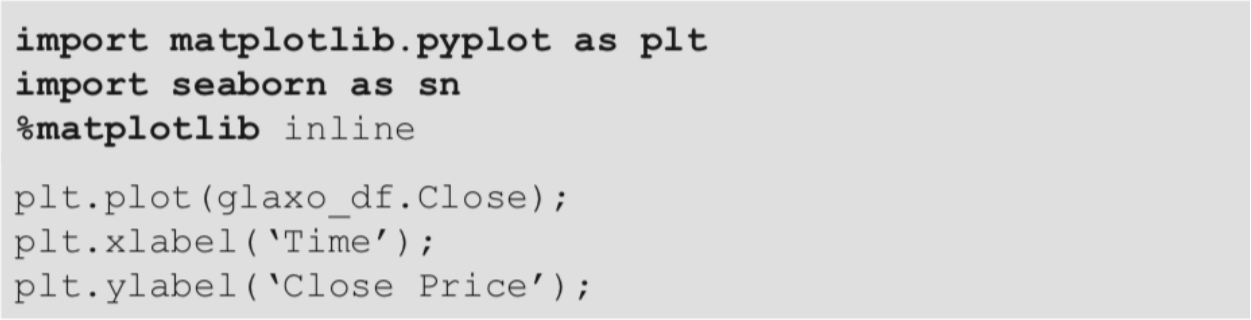


```
glaxo_df = glaxo_df.set_index(pd.DatetimeIndex(glaxo_df[‘Date’]))
beml_df = beml_df.set_index(pd.DatetimeIndex(beml_df[‘Date’]))
```

Normal Distribution (Cntd.)

Solution:

- Plotting the trend of close prices of GLAXO stock



```
import matplotlib.pyplot as plt
import seaborn as sn
%matplotlib inline

plt.plot(glaxo_df.Close);
plt.xlabel(‘Time’);
plt.ylabel(‘Close Price’);
```

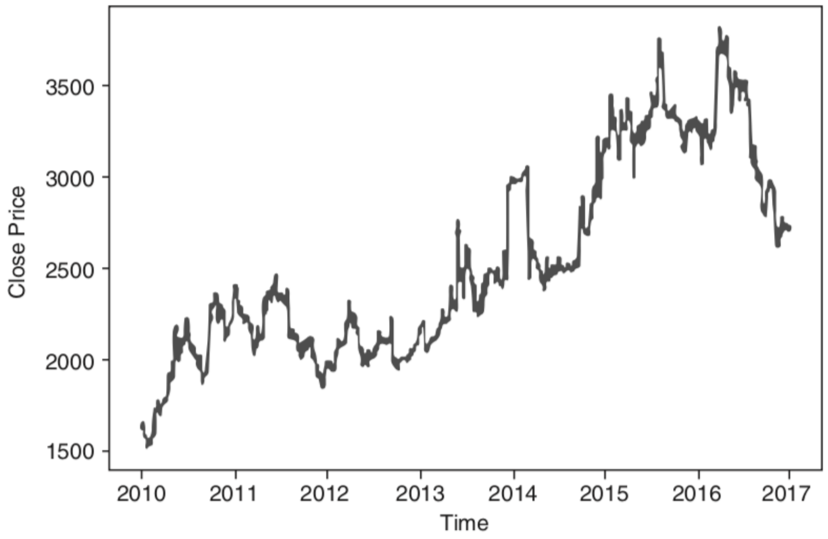
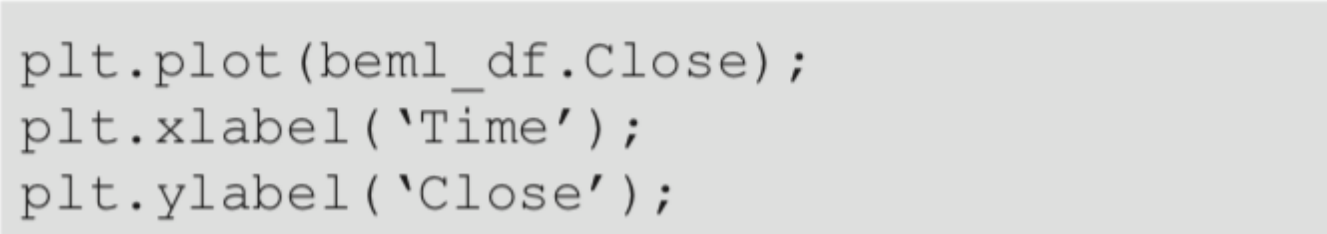


FIGURE 3.4 Close price trends of GLAXO stock.

Normal Distribution (Cntd.)

Solution:

- Plotting the trend of close prices of BEML stock



```
plt.plot(beml_df.Close);
plt.xlabel('Time');
plt.ylabel('Close');
```

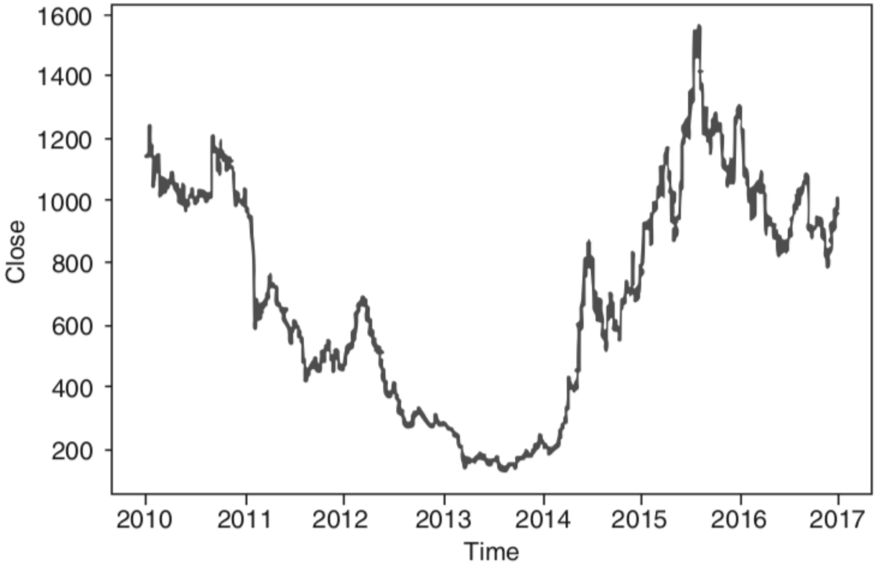


FIGURE 3.5 Close price trends of BEML stock.

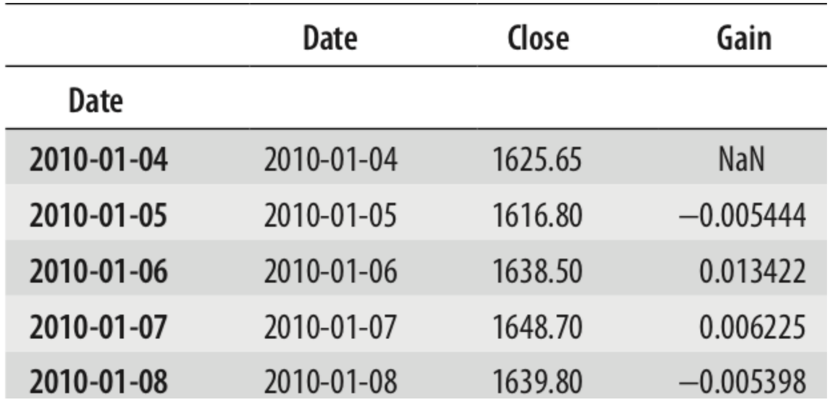
Normal Distribution (Cntd.)

- The behavior of daily returns on the stocks is called Gain.

$$gain = \frac{ClosePrice_t - ClosePrice_{t-1}}{ClosePrice_{t-1}}$$

- In Pandas it can be calculated as

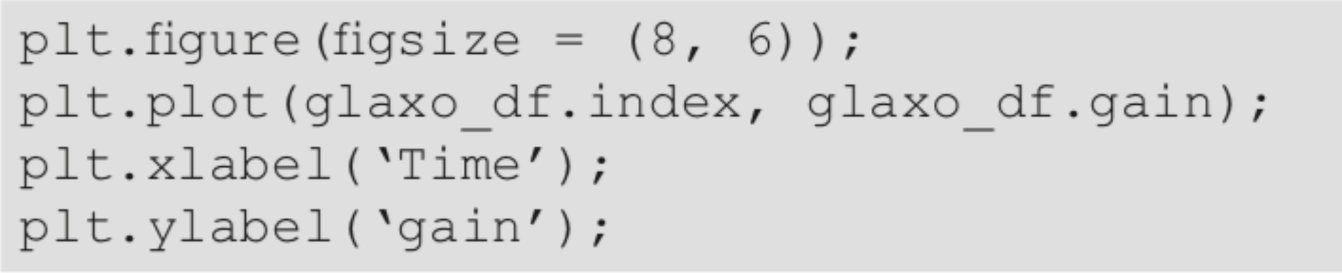
```
glaxo_df['gain'] = glaxo_df.Close.pct_change(periods = 1)
beml_df['gain'] = beml_df.Close.pct_change(periods = 1)
glaxo_df.head(5)
```



	Date	Close	Gain
Date			
2010-01-04	2010-01-04	1625.65	NaN
2010-01-05	2010-01-05	1616.80	−0.005444
2010-01-06	2010-01-06	1638.50	0.013422
2010-01-07	2010-01-07	1648.70	0.006225
2010-01-08	2010-01-08	1639.80	−0.005398

Normal Distribution (Cntd.)

- Plotting gain against time



```
plt.figure(figsize = (8, 6));
plt.plot(glaxo_df.index, glaxo_df.gain);
plt.xlabel('Time');
plt.ylabel('gain');
```

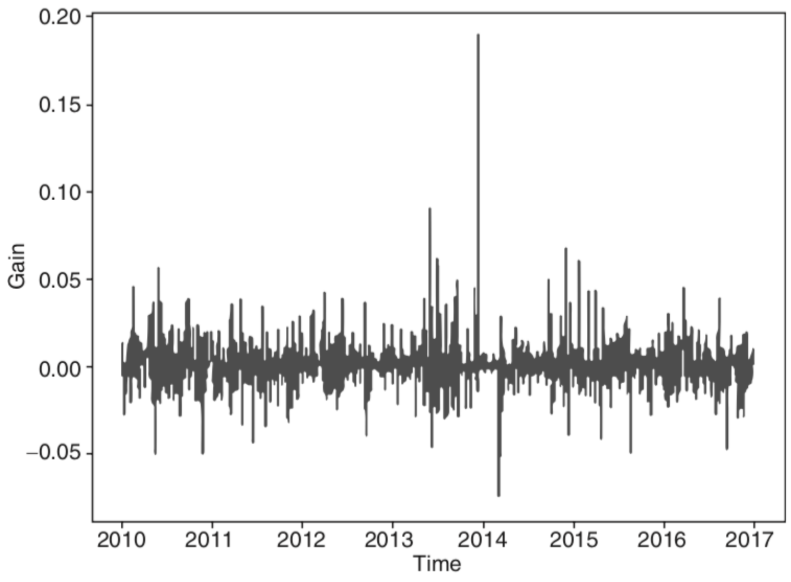
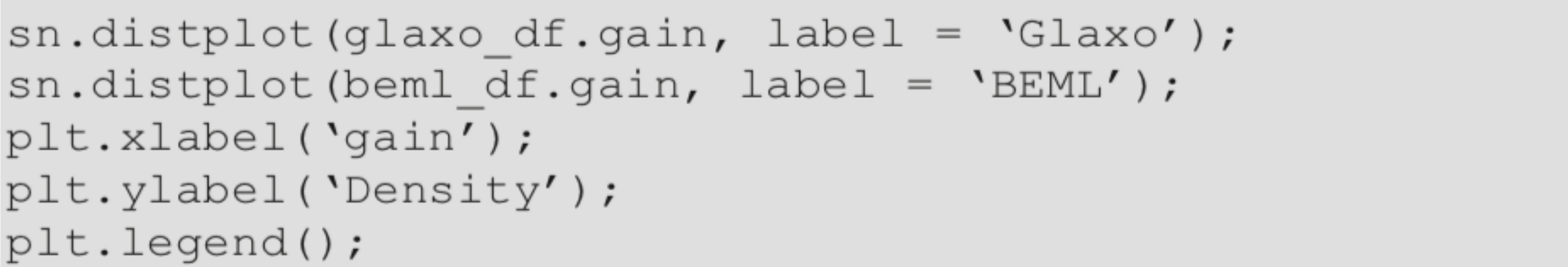


FIGURE 3.6 Daily gain of BEML stock.

Normal Distribution (Cntd.)

- Distribution plot of gain of both BEML and GLAXO stocks



```
sn.distplot(glaxo_df.gain, label = 'Glaxo');
sn.distplot(beml_df.gain, label = 'BEML');
plt.xlabel('gain');
plt.ylabel('Density');
plt.legend();
```

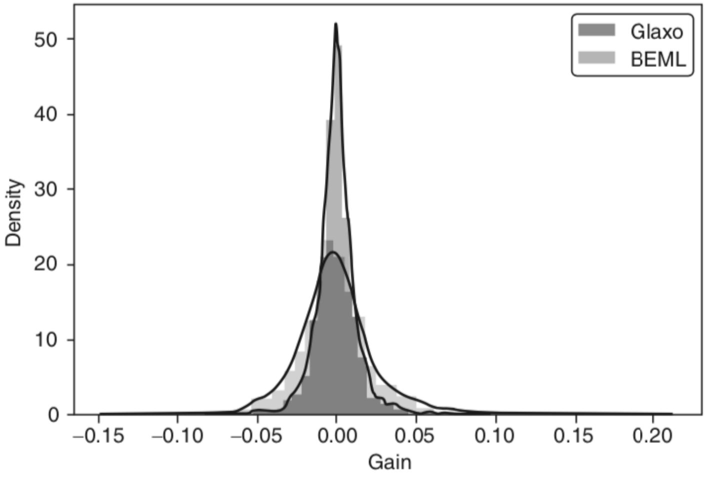


FIGURE 3.7 Distribution plot of daily gain of BEML and Glaxo stocks.

- Gain seems to be normally distributed for both the stocks with a mean around 0.00
- BEML seems to have a higher variance than GLAXO

Normal Distribution (Cntd.)

- The sample mean of a normal distribution is given by

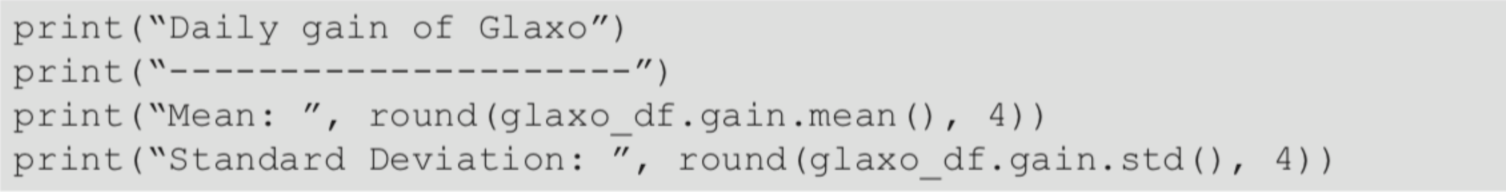
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Variance is given by

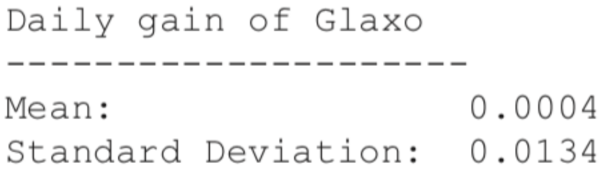
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Normal Distribution (Cntd.)

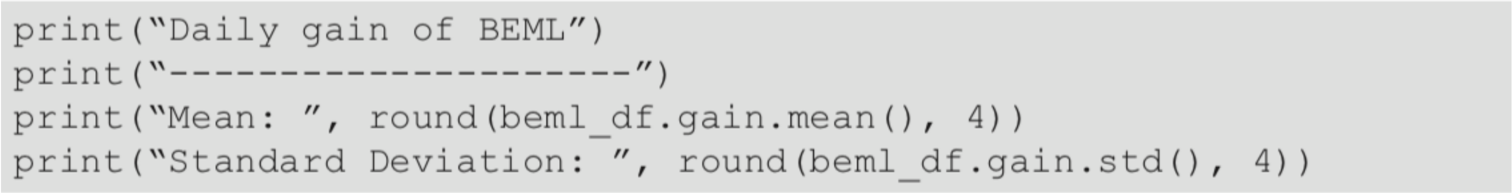
- In Pandas, the sample mean and standard deviation for daily returns for GLAXO and BEML are



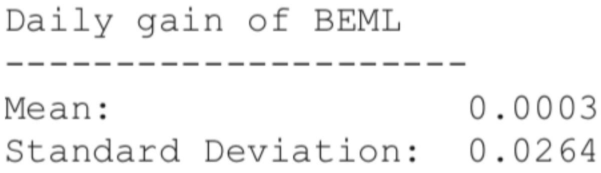
```
print("Daily gain of Glaxo")
print("---------------------")
print("Mean: ", round(glaxo_df.gain.mean(), 4))
print("Standard Deviation: ", round(glaxo_df.gain.std(), 4))
```



```
Daily gain of Glaxo
---------------------
Mean:                0.0004
Standard Deviation:  0.0134
```



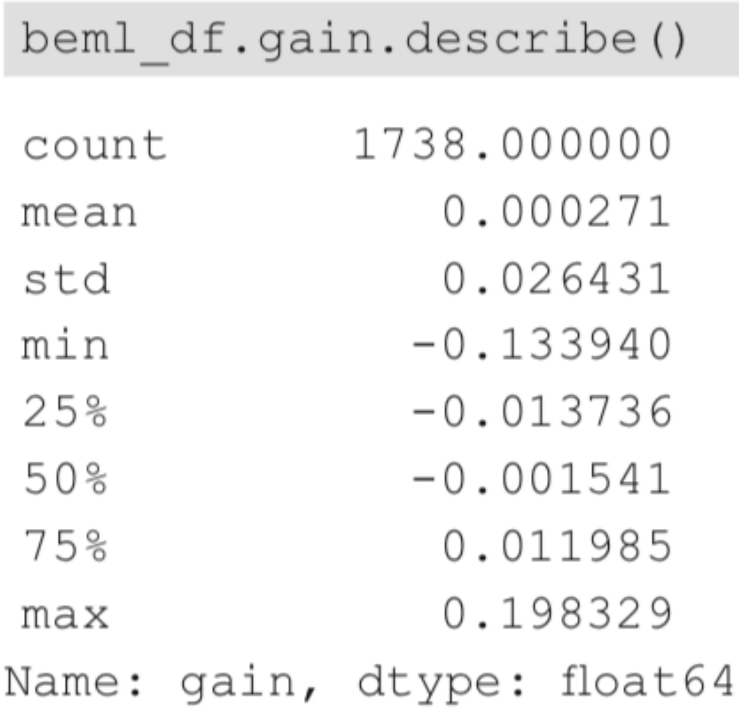
```
print("Daily gain of BEML")
print("---------------------")
print("Mean: ", round(beml_df.gain.mean(), 4))
print("Standard Deviation: ", round(beml_df.gain.std(), 4))
```



```
Daily gain of BEML
---------------------
Mean:                0.0003
Standard Deviation:  0.0264
```

Normal Distribution (Cntd.)

- The describe() method of DataFrame returns the detailed statistical summary of a variable



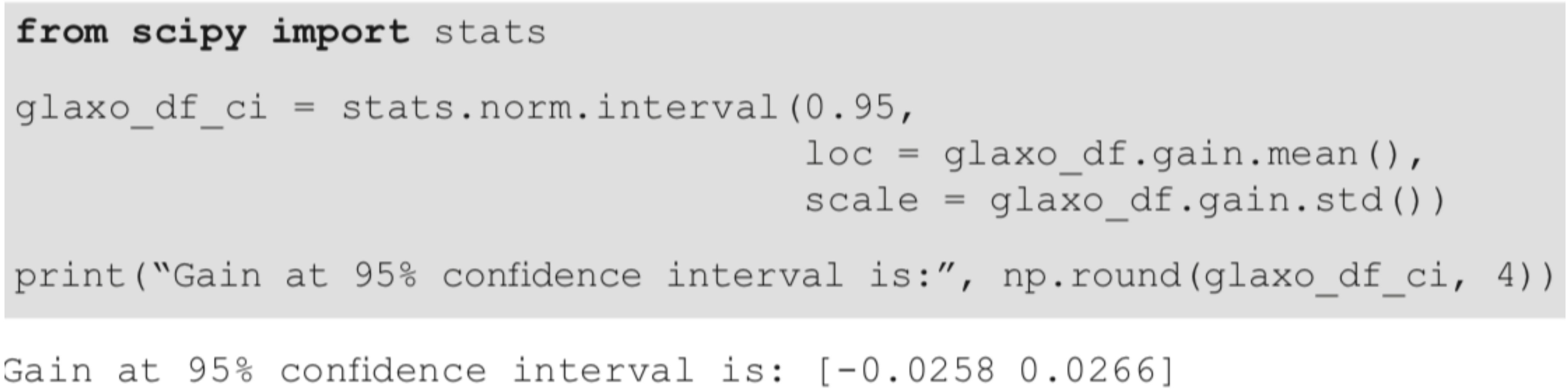
```
beml_df.gain.describe()
```

```
count      1738.000000
mean          0.000271
std           0.026431
min          -0.133940
25%          -0.013736
50%          -0.001541
75%           0.011985
max           0.198329
Name: gain, dtype: float64
```

- Expected daily rate of return (gain) is around 10% for both the stocks
- Variance or standard deviation of gain indicates risk.
- BEML stock has higher risk as standard deviation of BEML is 2.64% whereas the standard deviation for GLAXO is 1.33%

Normal Distribution (Cntd.)

- Gain at confidence interval 95% for a GLAXO stocks is given as



```
from scipy import stats

glaxo_df_ci = stats.norm.interval(0.95,
                                  loc = glaxo_df.gain.mean(),
                                  scale = glaxo_df.gain.std())

print(“Gain at 95% confidence interval is:”, np.round(glaxo_df_ci, 4))
```

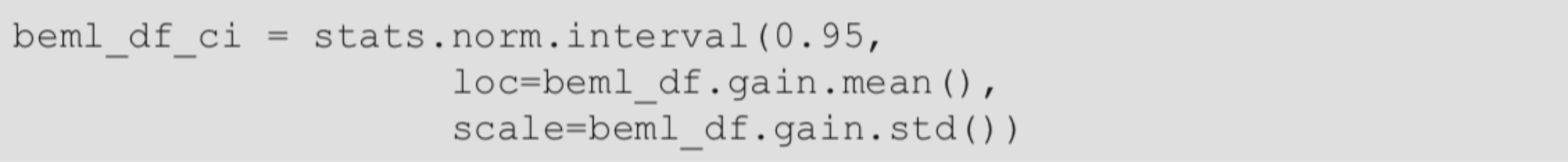
```
Gain at 95% confidence interval is: [-0.0258 0.0266]
```

Stats.norm.interval() takes three parameters

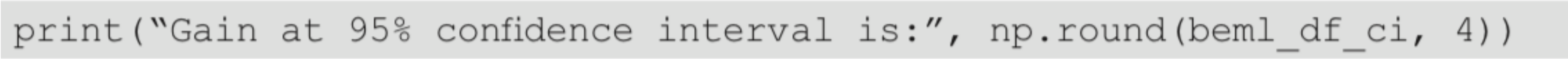
- **Alpha:** it is the interval
- **Loc:** location parameter, i.e. mean for normal distribution
- **Scale:** Scale parameter, i.e. standard deviation for normal distribution.

Normal Distribution (Cntd.)

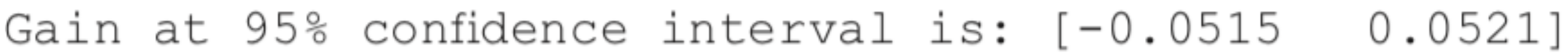
- Gain at confidence interval 95% for a GLAXO stocks is given as



```
beml_df_ci = stats.norm.interval(0.95,
                     loc=beml_df.gain.mean(),
                     scale=beml_df.gain.std())
```



```
print("Gain at 95% confidence interval is:", np.round(beml_df_ci, 4))
```



```
Gain at 95% confidence interval is: [-0.0515   0.0521]
```

- For 95% confidence interval, gain of GLAXO remains between -2.58% and 2.66% whereas gain of BEML remains between -5.15% and 5.21%.

Normal Distribution (Cntd.)

- **Cumulative Distribution Function** : Cumulative distribution function F(a) is the area under the probability density function up to X=a.
- The cumulative density function of a continuous random variable is shown below

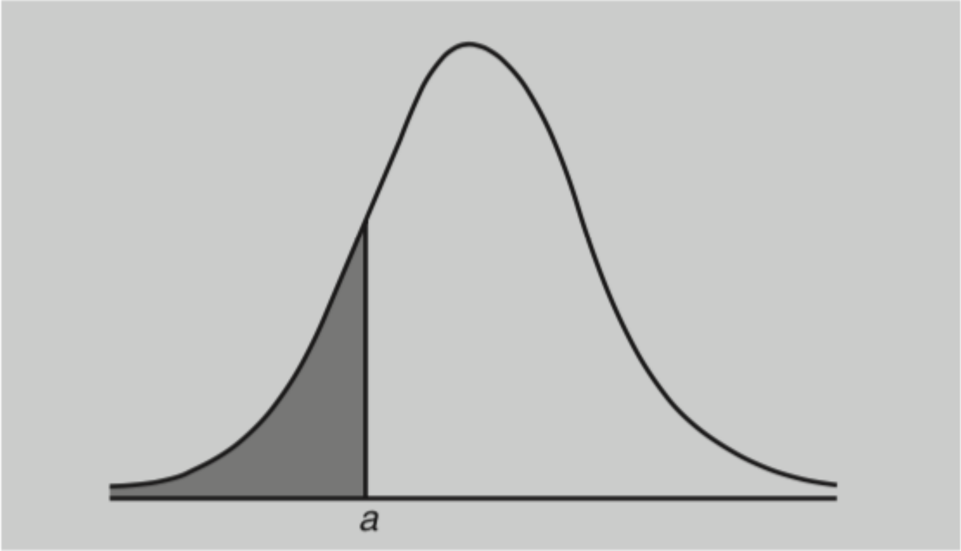
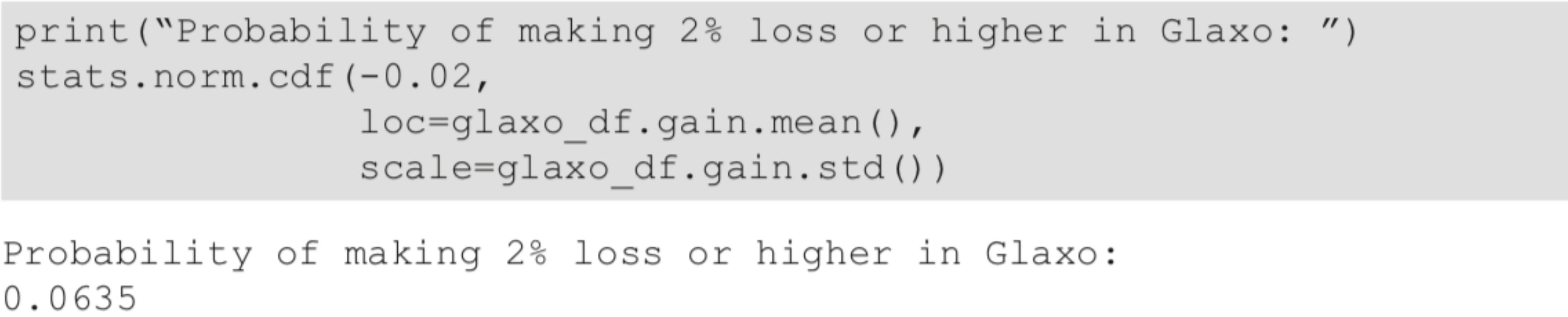


FIGURE 3.8 Cumulative distribution function *F*(*a*).

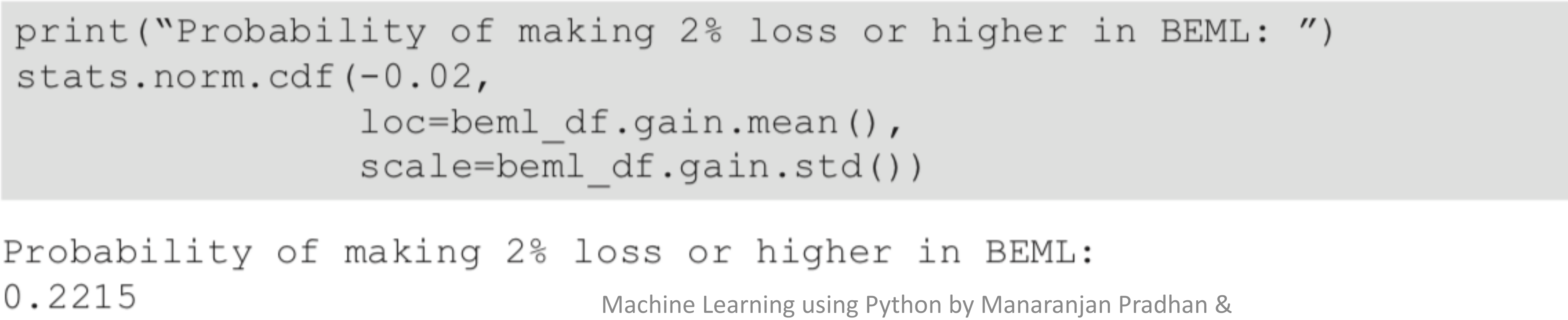
Normal Distribution (Cntd.)

The stats.norm.cdf() class returns cumulative distribution for a normal distribution.



```
print(“Probability of making 2% loss or higher in Glaxo: ”)
stats.norm.cdf(-0.02,
               loc=glaxo_df.gain.mean(),
               scale=glaxo_df.gain.std())
```

```
Probability of making 2% loss or higher in Glaxo:
0.0635
```

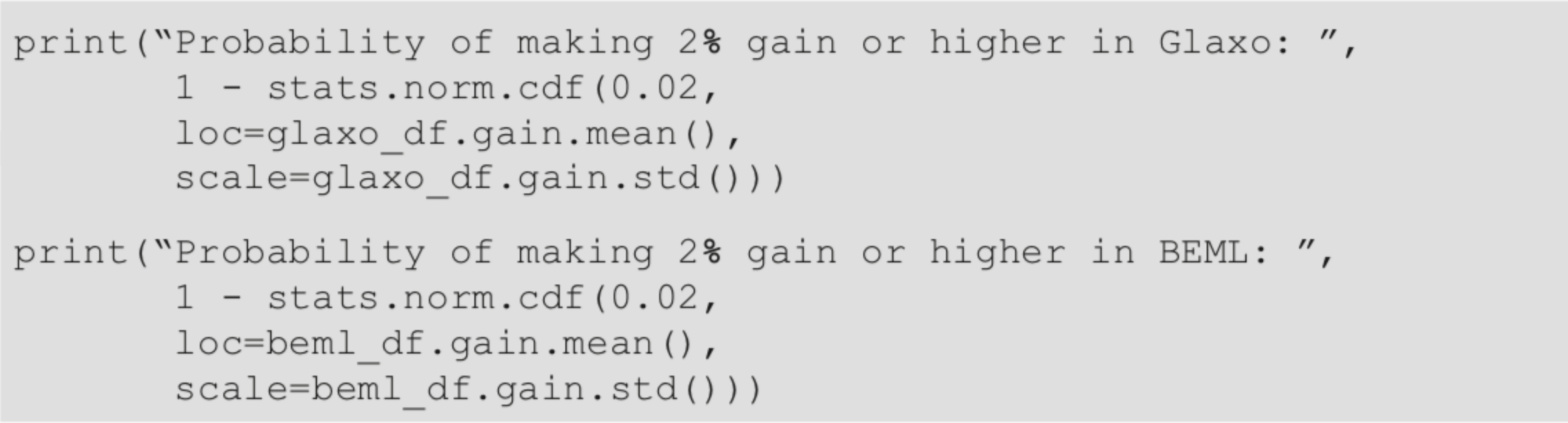


```
print(“Probability of making 2% loss or higher in BEML: ”)
stats.norm.cdf(-0.02,
               loc=beml_df.gain.mean(),
               scale=beml_df.gain.std())
```

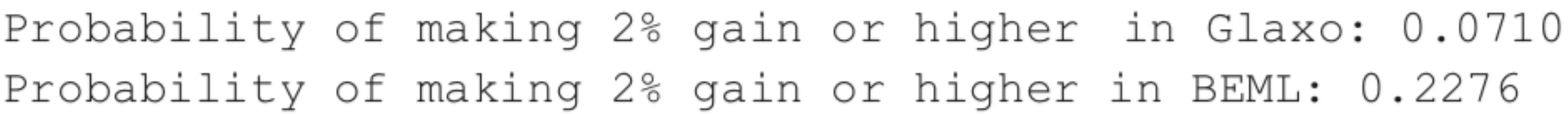
```
Probability of making 2% loss or higher in BEML:
0.2215
```

Normal Distribution (Cntd.)

The probability of making a daily gain of 2% or higher will be given by the area to the right of 0.02 of the distribution.



```
print(“Probability of making 2% gain or higher in Glaxo: ”,
       1 - stats.norm.cdf(0.02,
       loc=glaxo_df.gain.mean(),
       scale=glaxo_df.gain.std()))

print(“Probability of making 2% gain or higher in BEML: ”,
       1 - stats.norm.cdf(0.02,
       loc=beml_df.gain.mean(),
       scale=beml_df.gain.std()))
```



```
Probability of making 2% gain or higher  in Glaxo: 0.0710
Probability of making 2% gain or higher in BEML: 0.2276
```

Other Important Distributions

1. **Beta Distribution:** It is a continuous distribution which can take values between 0 and 1.
2. **Gamma Distribution:** It is used to model waiting times in different contexts, We can model time until the next n events occur using Gamma distribution.
3. **Weibull Distribution:** Exponential distribution is a special case of Weibull distribution. Using Weibull distribution we can model increasing (or decreasing) rates of occurrence of an event over time.
4. **Geometric distribution:** It is used to solve ”How many failures until a success?”

CENTRAL LIMIT THEOREM

- Let $S_{1,}$ $S_{2,\ldots}$ S_k be samples of size n drawn from an independent and identically distributed population with mean μ and standard deviation σ.
- Let $X_{1,}$ $X_{2,\ldots}$ X_k be the sample means.
- "According to the CLT, the distribution of $X_{1,}$ $X_{2,\ldots}$ X_k follows a normal distribution with mean μ and standard deviation of $\sigma/\sqrt{n}$. That is, the sampling distribution of mean will follow a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$."

HYPOTHESIS TEST

- Hypothesis is a claim
- Hypothesis testing is to either reject or retain a null hypothesis using data
- Hypothesis testing consists of two complementary statements
  1. Null Hypothesis (H_0)– an existing belief
  2. Alternate Hypothesis (H_A) – what we intend to establish with new evidences
- Hypothesis tests are broadly classified into
  1. Parametric tests – about population parameters such as mean, proportion, standard deviations etc.
  2. Non-parametric tests – about non parametric characteristics such as independence of events or data following certain distributions such as normal distribution.

HYPOTHESIS TEST (Cntd.)

- Few examples of the null hypothesis are as follows:
  1. Children who drink the health drink Complan are likely to grow taller
  2. Women use camera phone more than men (Freier, 2016)
  3. Vegetarians miss few flights (Siegel, 2016)
  4. Smokers are better sales people

HYPOTHESIS TEST (Cntd.)

- The steps for hypothesis tests are as follows:
  1. Define null and alternate hypothesis. Hypothesis is described using a population parameter.
  2. Identify the test statistic to be used for testing the validity of the null hypothesis
  3. Decide the criteria for rejection and retention of null hypothesis. This is called significance value.
  4. Calculate the p-value (probability value), which is the conditional probability of observing the test statistic value when the null hypothesis is true.
  5. Take the decision to reject or retain the null hypothesis based on the p-value and the significance value.

HYPOTHESIS TEST (Cntd.)

- For all the following examples, we will use the following notations:
  1. μ - population mean
  2. σ - population standard deviation
  3. X - sample mean
  4. S - sample standard deviation
  5. n - sample size

Z-test

- Z-test is used when:
    1. We need to test the value of population mean, given that population variance is known
    2. The population is a normal distribution and the population variance is known
    3. The sample size is large and the population variance is known. That is, the assumption of normal distribution can be relaxed for large samples (n>30)
- Z- statistic is calculated as

$$Z = \frac{X - \mu}{\sigma / \sqrt{n}}$$

Z-test (Cntd.)

Example:

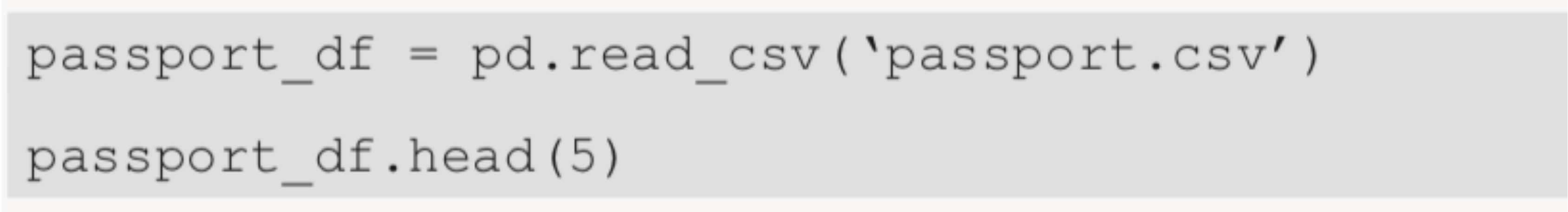
A passport office claims that the passport applications are processed within 30 days of submitting the application form and all necessary documents. The file *passport.csv* contains processing time of 40 passport applicants. The population standard deviation of the processing time is 12.5 days. Conduct a hypothesis test at significance level ∝ =0.05 to verify the claim made by the passport office.

In this case, the population mean (claim made by passport office) and standard deviation are known. Population mean is 30 and population standard deviation is 12.5. The dataset in *passport.csv* contains observations of actual processing time of 40 passports. We can calculate the mean of these observations and calculate Z-statistic less than -1.64, then it can be concluded that the processing time is less than 30 as claimed by passport office.

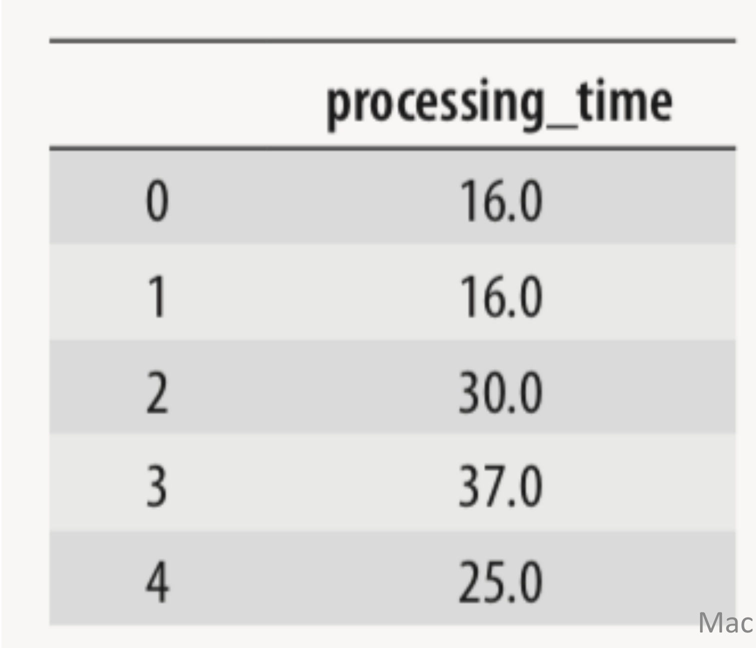
Z-test (Cntd.)

Solution:

Load the data and display first 5 records from *passport.csv*



```
passport_df = pd.read_csv(‘passport.csv’)

passport_df.head(5)
```

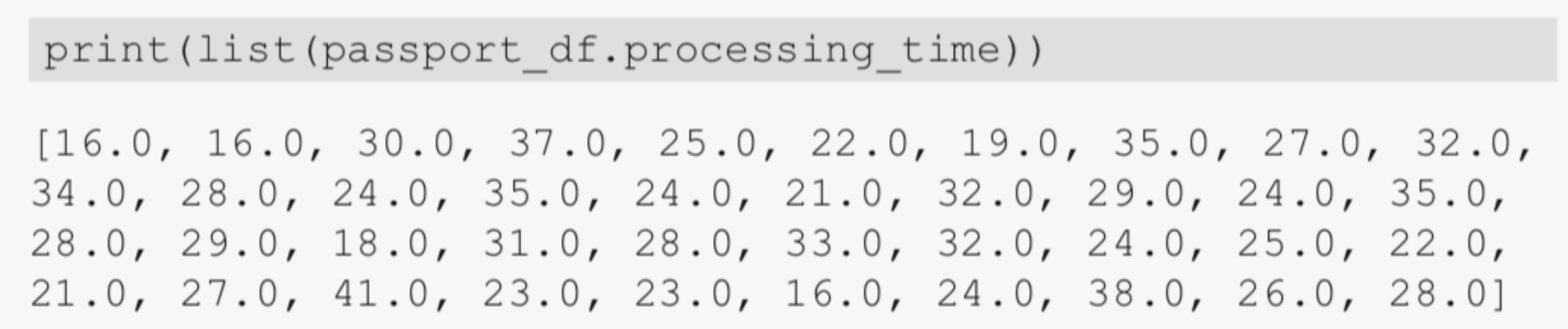


	processing_time
0	16.0
1	16.0
2	30.0
3	37.0
4	25.0

Z-test (Cntd.)

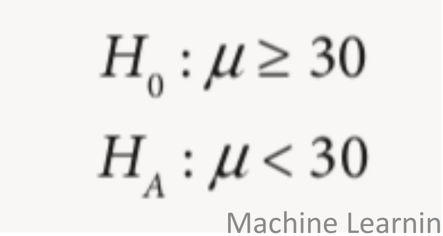
Solution :

Printing all the records from the dataset



```
print(list(passport_df.processing_time))

[16.0, 16.0, 30.0, 37.0, 25.0, 22.0, 19.0, 35.0, 27.0, 32.0,
34.0, 28.0, 24.0, 35.0, 24.0, 21.0, 32.0, 29.0, 24.0, 35.0,
28.0, 29.0, 18.0, 31.0, 28.0, 33.0, 32.0, 24.0, 25.0, 22.0,
21.0, 27.0, 41.0, 23.0, 23.0, 16.0, 24.0, 38.0, 26.0, 28.0]
```

Defining the hypothesis

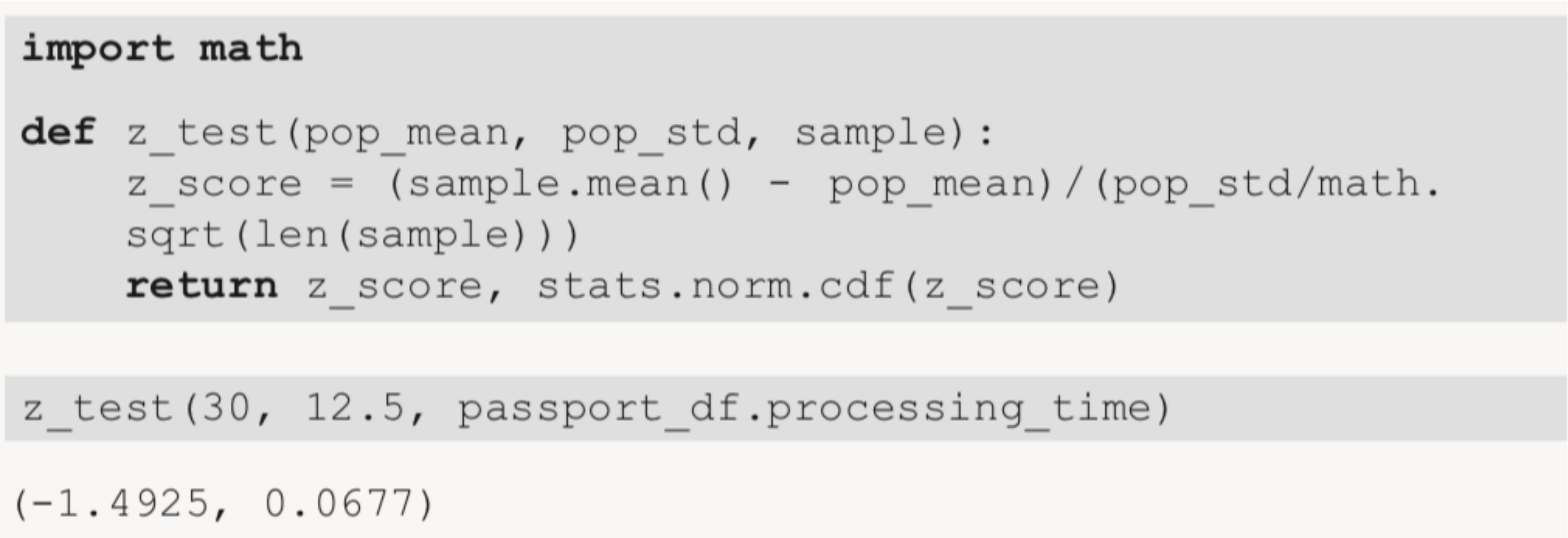


$$H_0 : \mu \geq 30$$

$$H_A : \mu < 30$$

Z-test (Cntd.)
Solution :

- Conducting Z-test for the above hypothesis test



```
import math

def z_test(pop_mean, pop_std, sample):
    z_score = (sample.mean() - pop_mean)/(pop_std/math.
    sqrt(len(sample)))
    return z_score, stats.norm.cdf(z_score)
```

```
z_test(30, 12.5, passport_df.processing_time)
```

```
(-1.4925, 0.0677)
```

- The first value of the result is Z-statistic value or Z-score and second value is the corresponding p-value.

Z-test (Cntd.)
Solution :

Results:

- As the p-value is more than 0.05, and Z-statistics value is higher than -1.64
- There is 6.77% probability of observing a random sample at least as extreme as the observed sample.
- Since 6.77% is greater than the significance value 5%, there is not enough evidence to reject null hypothesis.
- Hence, the null hypothesis is retained.

One-Sample t-Test

- t-test is used when the population standard deviation S is unknown (and hence estimated from the sample)
- It is estimated from the sample.
- Mathematically

$$t\text{-Statistics} = \frac{(\bar{X} - \mu)}{S/\sqrt{n}}$$

- The expected value of a sample of independent observations is equal to the given population mean.

One-Sample t-Test (Cntd.)

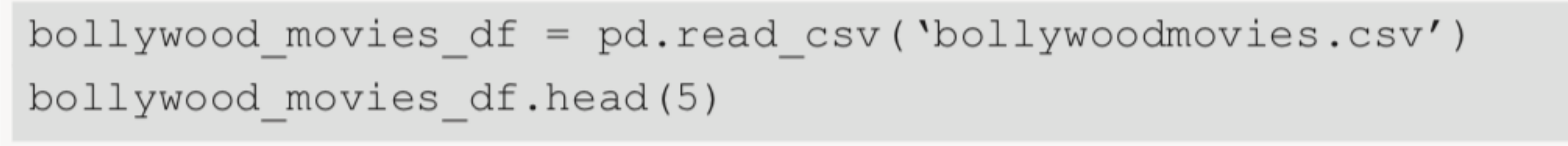
Example:

Aravind Productions (AP) is a newly formed movie production house based out of Mumbai, India. AP was interested in understanding the production cost required for producing Bollywood movie. The industry believes that the production house will require INR 500 million on average. It is assumed that the Bollywood movie production cost follows a normal distribution. The production costs of 40 Bollywood movies in millions of rupees are given in *bollywoodmovies.csv* file. Conduct and appropriate hypothesis test at $\propto$ = 0.05 to check whether the belief about average production cost is correct.

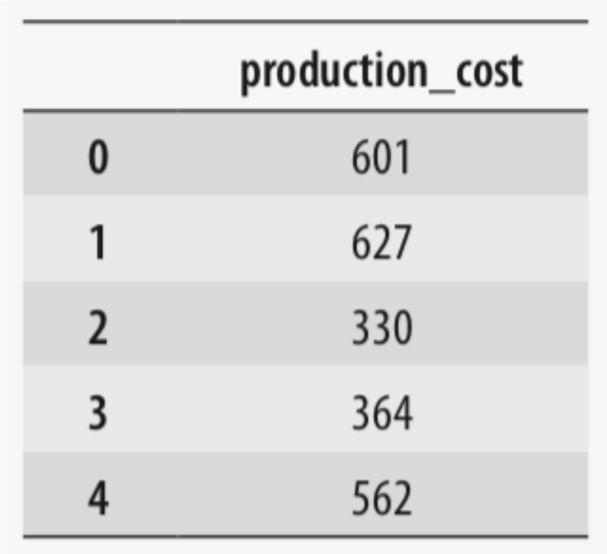
The population mean is 500 and the sample set for actual production cost is available in the file *bollywoodmovies.csv.* The population standard deviation is not known.

One-Sample t-Test (Cntd.)
Solution:

- Reading the data



```
bollywood_movies_df = pd.read_csv(‘bollywoodmovies.csv’)
bollywood_movies_df.head(5)
```

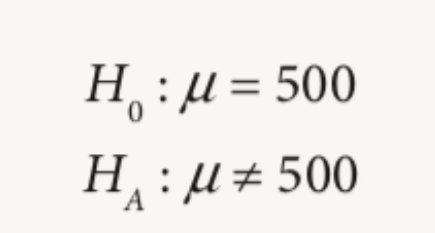


	production_cost
0	601
1	627
2	330
3	364
4	562

One-Sample t-Test (Cntd.)

Solution:

- Defining the hypothesis



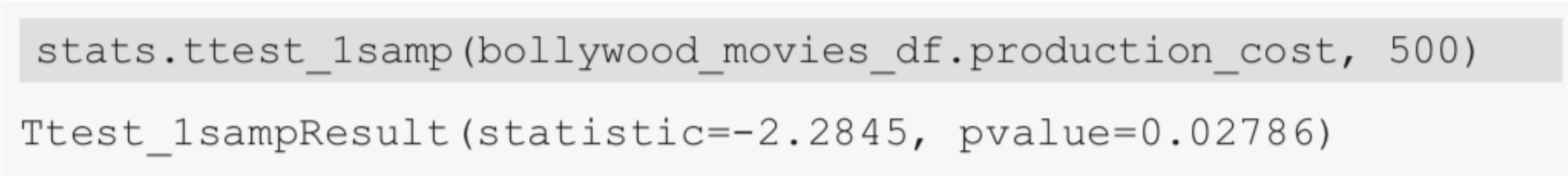
$$H_0 : \mu = 500$$
$$H_A : \mu \neq 500$$

- It takes two parameters

1. **a: array_like** – sample observation
2. **popmean: float –** expected value in null hypothesis

One-Sample t-Test (Cntd.)
Solution:

- Conducting the test



```
stats.ttest_1samp(bollywood_movies_df.production_cost, 500)

Ttest_1sampResult(statistic=-2.2845, pvalue=0.02786)
```

- t-statistic value = -2.2845, and p-value = 0.02786
- The sample mean is less than population mean
- p-value is less than 0.05
- Reject the null hypothesis

Two-Sample t-Test

- To test difference between two population means where standard deviations are unknown.

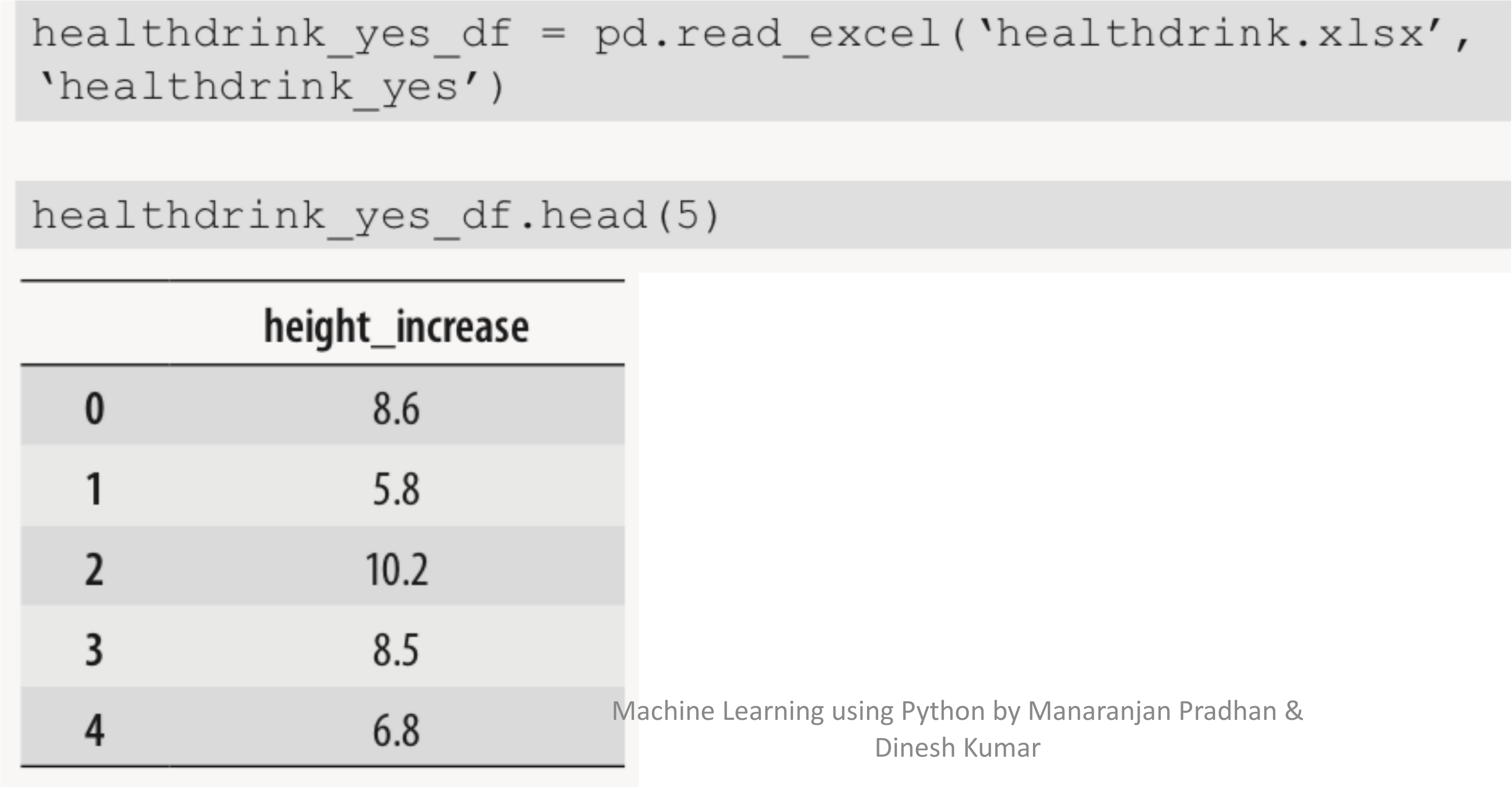
Example:

A company claims that children who drink their health drink will grow taller than the children who do not drink that health drink. Data in the file *healthdrink.xlsx* shows average increase in height over one-year period from two groups: one drinking the health drink and the other not drinking the health drink. At $\propto$ =0.05, test whether the increase in height for the children who drink the health drink is different than those who do not drink health drink.

Two-Sample t-Test (Cntd.)

Solution:

- Reading the data with the tab *healthdrink_yes* as parameter and then display first five records



```
healthdrink_yes_df = pd.read_excel('healthdrink.xlsx',
'healthdrink_yes')
```

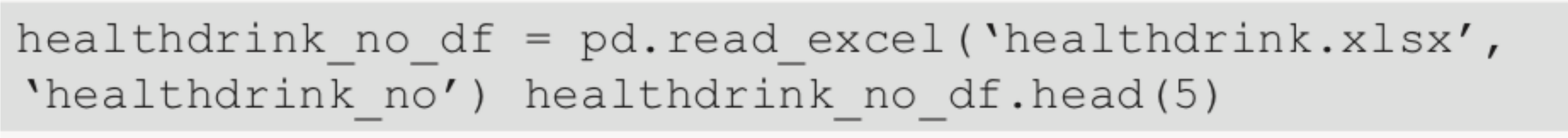
```
healthdrink_yes_df.head(5)
```

	height_increase
0	8.6
1	5.8
2	10.2
3	8.5
4	6.8

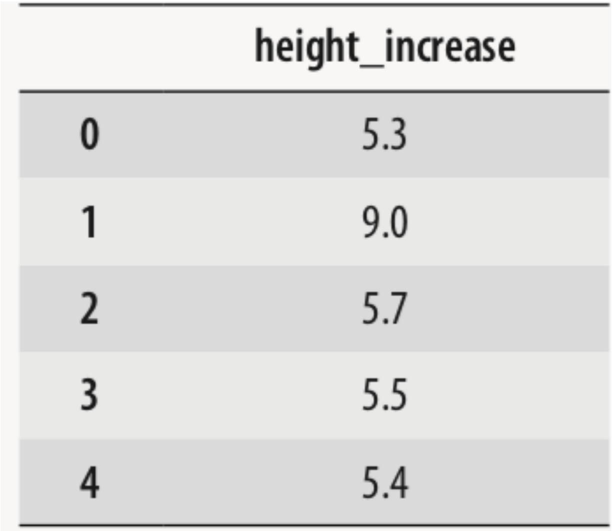
Two-Sample t-Test (Cntd.)

Solution:

- Reading the data from the tab *healthdrink_no* in data and then display first five records



```
healthdrink_no_df = pd.read_excel('healthdrink.xlsx',
'healthdrink_no') healthdrink_no_df.head(5)
```



	height_increase
0	5.3
1	9.0
2	5.7
3	5.5
4	5.4

Two-Sample t-Test (Cntd.)

Solution:

- The distribution plots of increase in height separately for drinking health drink and not drinking health drink groups

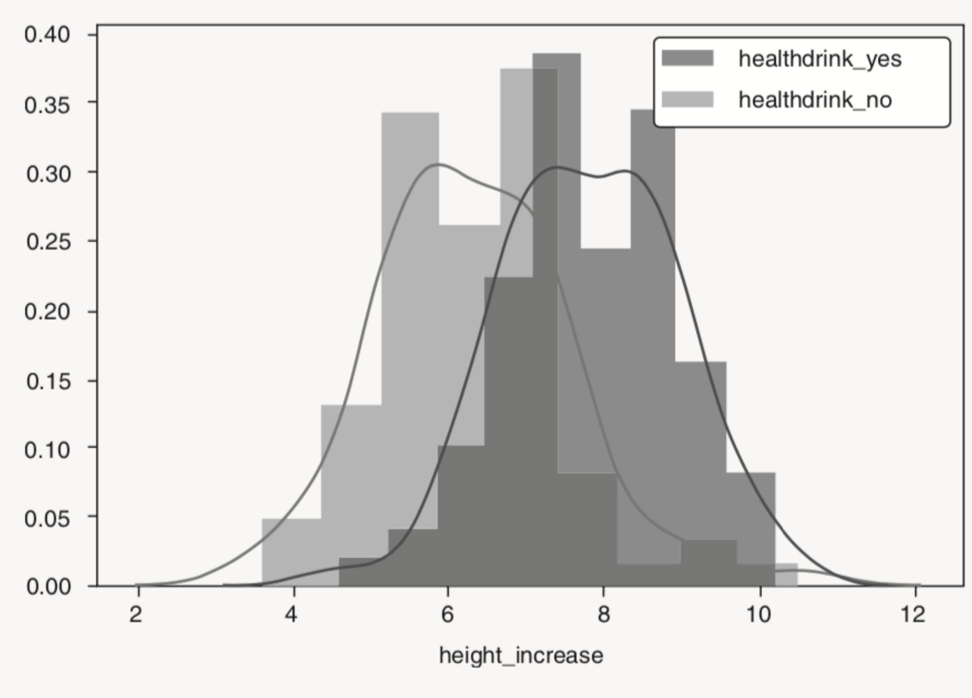


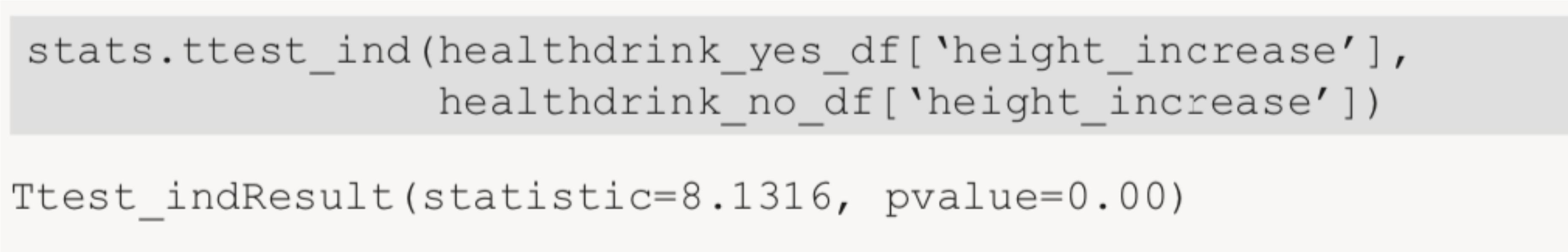
FIGURE 3.9 Comparing distributions for health drink and no health drink.

- The distribution of increase in height for those who have the health drink has shifted to the right of those who did not have the health drink.
- But is the difference, as claimed, statistically significant?

Two-Sample t-Test (Cntd.)

Solution:

- Conducting the test



```
stats.ttest_ind(healthdrink_yes_df[‘height_increase’],
                healthdrink_no_df[‘height_increase’])
```

```
Ttest_indResult(statistic=8.1316, pvalue=0.00)
```

- Probability of the samples belonging to the same distribution is almost 0.
- This means that the increase in height for those who had health drink is significantly different than those who did not.

Paired-Sample t-Test

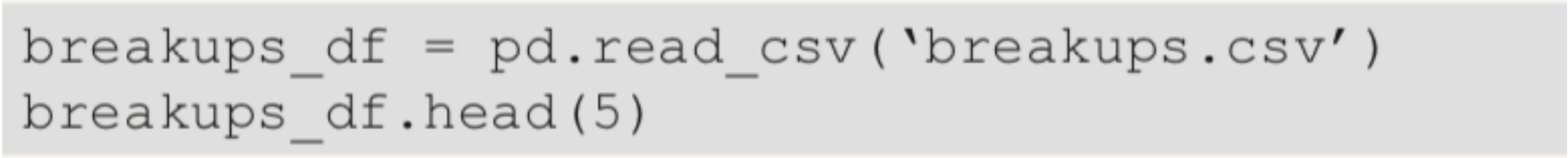
- When we need to check whether the difference in the parameter values is statistically significant before and after the intervention or between two different types of interventions.
- Used for comparing two different interventions applied on the same sample.
- For example
  1. Performance of employees after a training program
  2. Treatment of specific illness

Paired-Sample t-Test (Cntd.)

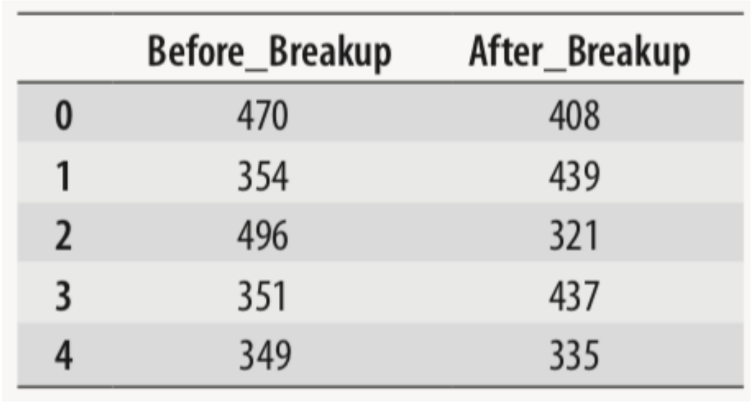
Example:

The file *breakup.csv* contains alcohol consumption before and after breakup. Conduct a paired t-test to check whether the alcohol consumption is more after the breakup at 95% confidence.

Solution: Reading the data and displaying first 5 records



```
breakups_df = pd.read_csv('breakups.csv')
breakups_df.head(5)
```

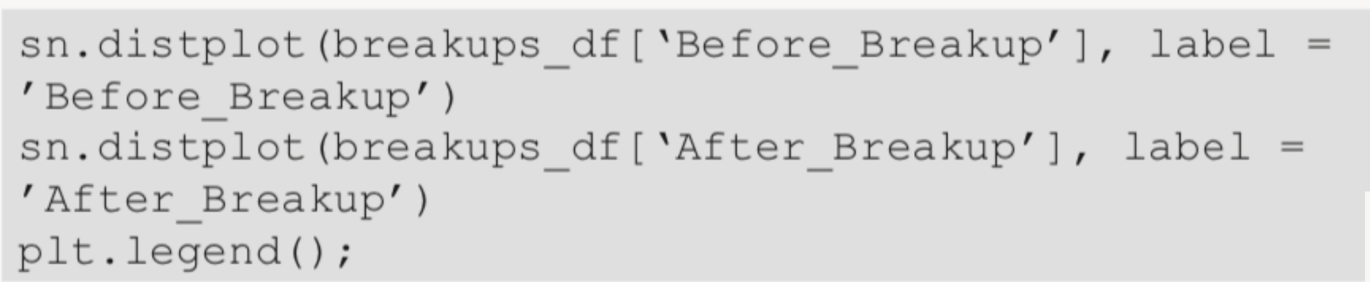


	Before_Breakup	After_Breakup
0	470	408
1	354	439
2	496	321
3	351	437
4	349	335

Paired-Sample t-Test (Cntd.)

Solution:

The distribution plots of alcohol consumption separately for before and after breakups.



```
sn.distplot(breakups_df[‘Before_Breakup’], label =
’Before_Breakup’)
sn.distplot(breakups_df[‘After_Breakup’], label =
’After_Breakup’)
plt.legend();
```

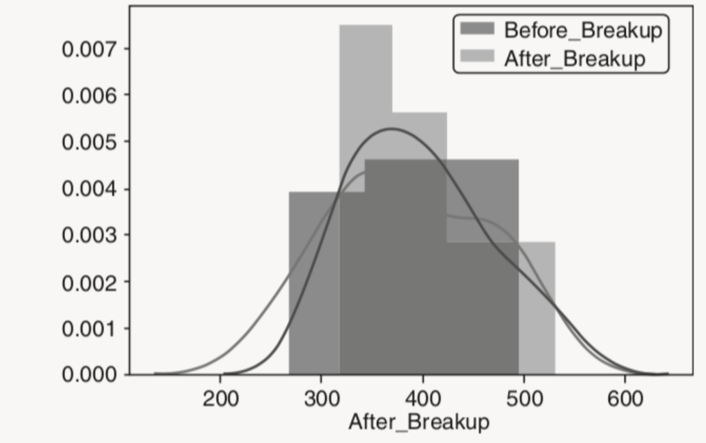
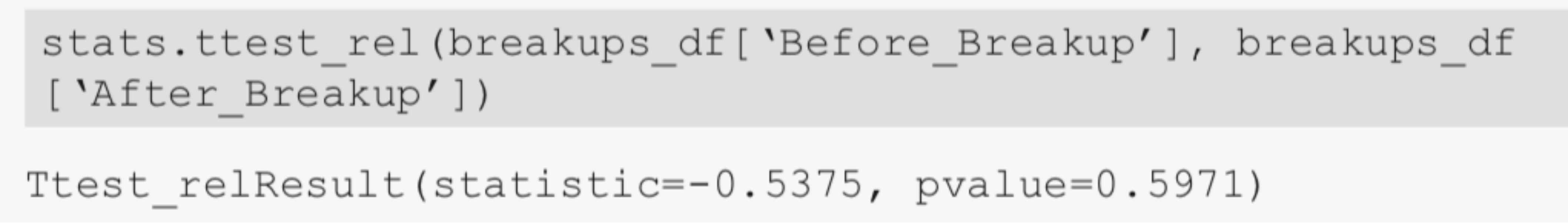


FIGURE 3.10 Comparing distributions for drinking habits before and after breakups.

Paired-Sample t-Test (Cntd.)

Solution:

- Conducting the test



```
stats.ttest_rel(breakups_df[‘Before_Breakup’], breakups_df
[‘After_Breakup’])

Ttest_relResult(statistic=-0.5375, pvalue=0.5971)
```

- Probability of the samples belonging to the same distribution is 0.597 which is more than 0.05 value.
- We conclude that they are part of same distribution
- There is no change in alcohol consumption pattern before and after breakup

Chi-Square Goodness of Fit Test

- A non-parametric test use for comparing the observed distribution of data with the expected distribution of the data
- Used to decide whether there is any statistically significant difference between the observed distribution and a theoretical distribution.
- Chi-square statistics is given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

- Where O_i is the observed frequency
- E_i is the expected frequency of the i^{th} category.

Chi-Square Goodness of Fit Test (Cntd.)

Example:

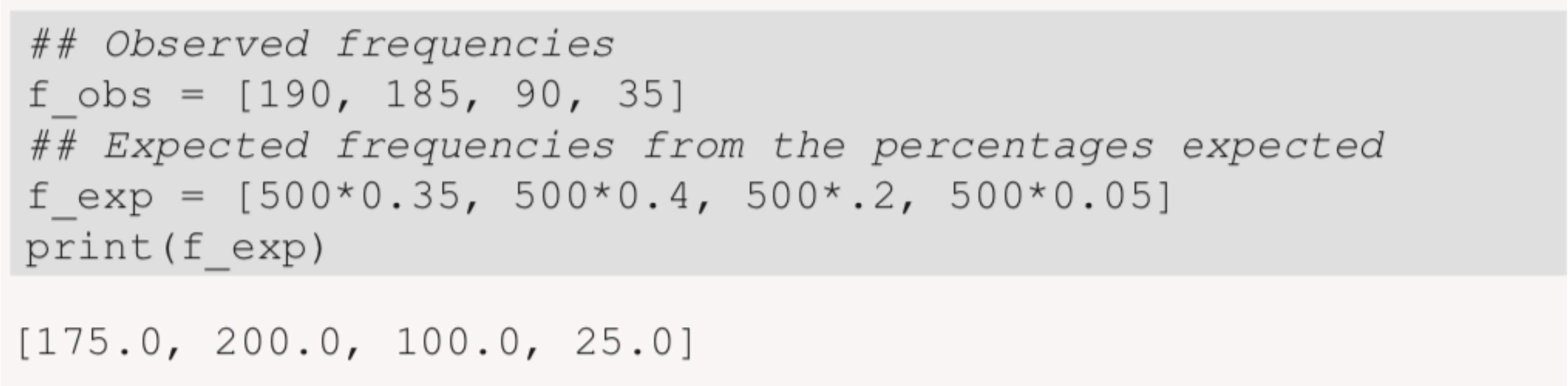
Hanuman Airlines (HA) operated daily flights to several Indian cities. One of the problems HA faces is the food preferences by the passengers. Captain Cook, the operations manager of HA, believes that 35% of their passengers prefer vegetarian food, 40% prefer non-vegetarian food, 20% low calorie food, and 5% request for diabetic food. A sample of 500 passengers was chosen to analyze the food preferences and the observed frequencies are as follows:

1. Vegetarian: 190
2. Non-vegetarian: 185
3. Low-calorie: 90
4. Diabetic: 35

Conduct a chi-square test to check whether Captain Cook’s belief is true at $\propto$=0.05.

Chi-Square Goodness of Fit Test (Cntd.)
Solution:

- Parameters used for chi-square test
  1. f_obs: array_like – Observed frequencies in each category
  2. f_exp: array_like – Expected frequencies in each category
- From the data we can create the following arrays:



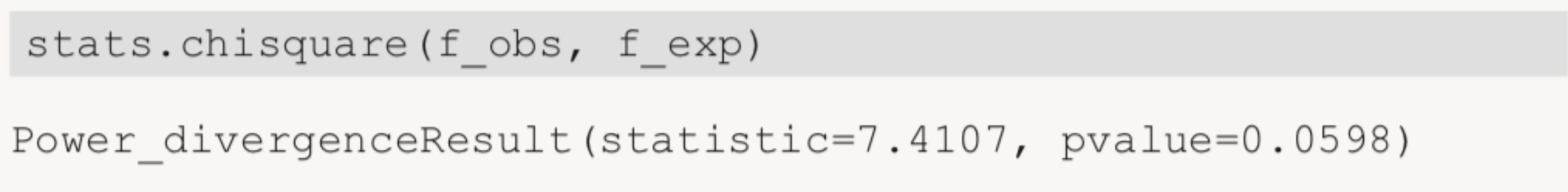
```
## Observed frequencies
f_obs = [190, 185, 90, 35]
## Expected frequencies from the percentages expected
f_exp = [500*0.35, 500*0.4, 500*.2, 500*0.05]
print(f_exp)
```

```
[175.0, 200.0, 100.0, 25.0]
```

Chi-Square Goodness of Fit Test (Cntd.)

Solution:

- Conducting the test



```
stats.chisquare(f_obs, f_exp)
```

```
Power_divergenceResult(statistic=7.4107, pvalue=0.0598)
```

- P-value is more than 0.05 value.
- We retain the null hypothesis.
- That is, Captain Cook’s belief is true.

ANALYSIS OF VARIANCE (ANOVA)

- **One-way ANOVA** – used to study the impact of a single treatment at different levels on a continuous response variable.
- The null and alternative hypothesis for comparing 3 groups are given by

$$H_0\text{: } \mu_1 = \mu_2 = \mu_3$$

$$H_A\text{: Not all } \mu \text{ values are equal}$$

- Note that the alternative hypothesis, not all values are equal, implies that some of groups could be equal.

ANALYSIS OF VARIANCE (ANOVA) (Cntd.)

Example:

Ms Rachael Khanna the brand manager of ENZO detergent powder at the "one-stop" retail was interested in understanding whether the price discounts have any impact on the sales quantity of ENZO. To test whether the price discounts had any impact, price discounts of 0%, 10%, and 20% were given on randomly selected days. The quantity of ENZO sold in a day under different discount levels is shown in Table 3.1. Conduct a one way ANOVA to check whether discount had any significant impact on the average sales quantity at $\propto = 0.05$

ANALYSIS OF VARIANCE (ANOVA) (Cntd.)

Example:

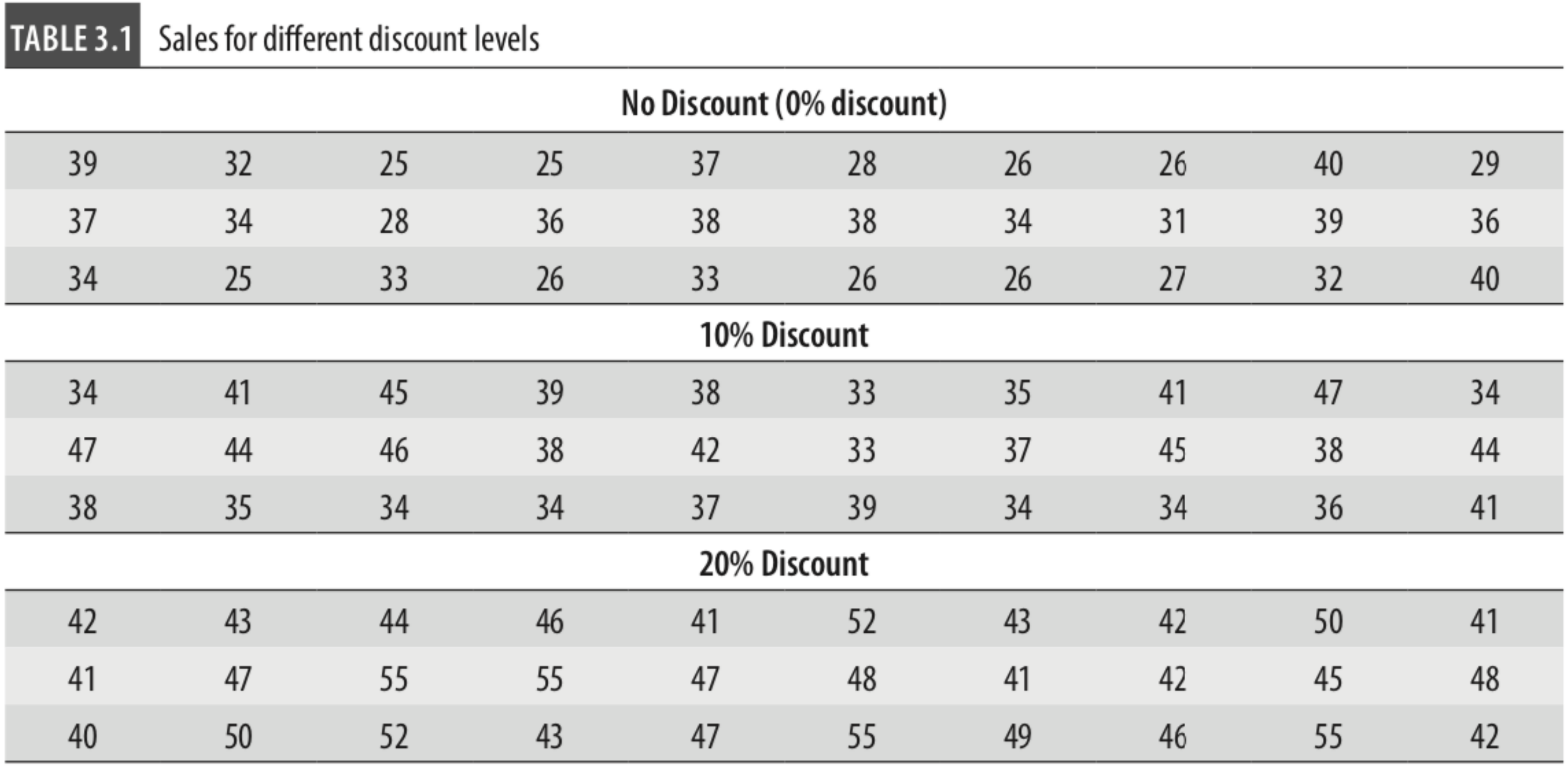


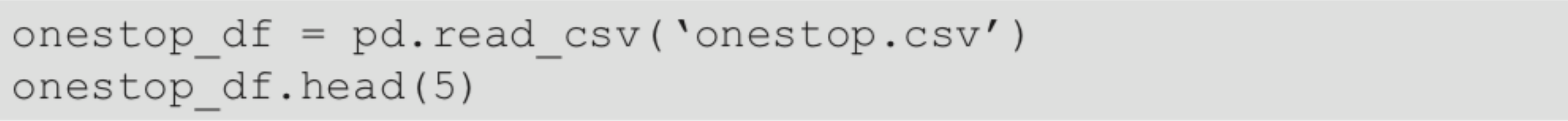
TABLE 3.1 Sales for different discount levels

No Discount (0% discount)									
39	32	25	25	37	28	26	26	40	29
37	34	28	36	38	38	34	31	39	36
34	25	33	26	33	26	26	27	32	40
10% Discount									
34	41	45	39	38	33	35	41	47	34
47	44	46	38	42	33	37	45	38	44
38	35	34	34	37	39	34	34	36	41
20% Discount									
42	43	44	46	41	52	43	42	50	41
41	47	55	55	47	48	41	42	45	48
40	50	52	43	47	55	49	46	55	42

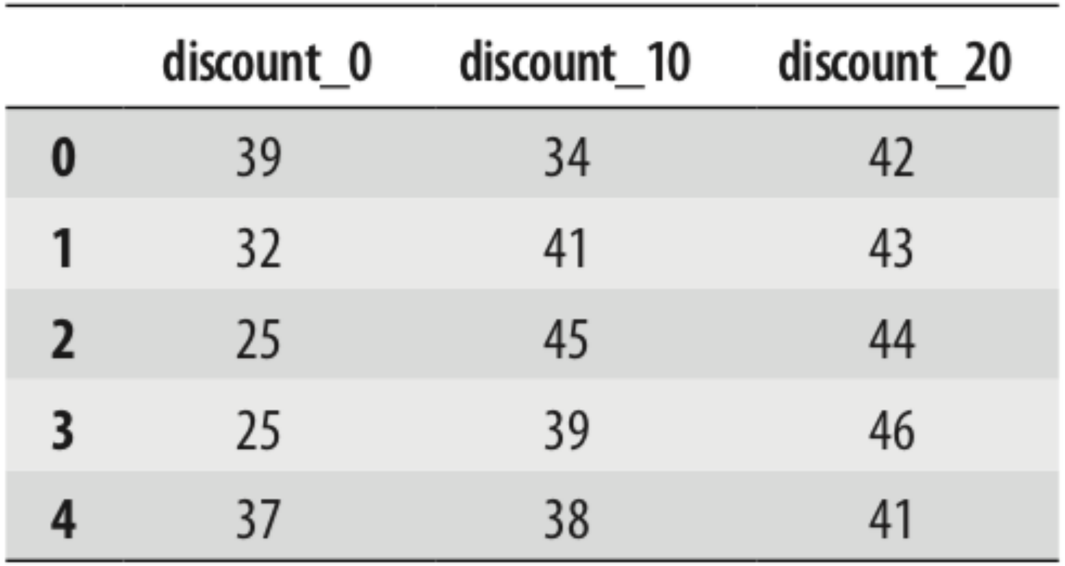
ANALYSIS OF VARIANCE (ANOVA) (Cntd.)

Solution:

Read the records from the file and print the first few records



```
onestop_df = pd.read_csv('onestop.csv')
onestop_df.head(5)
```

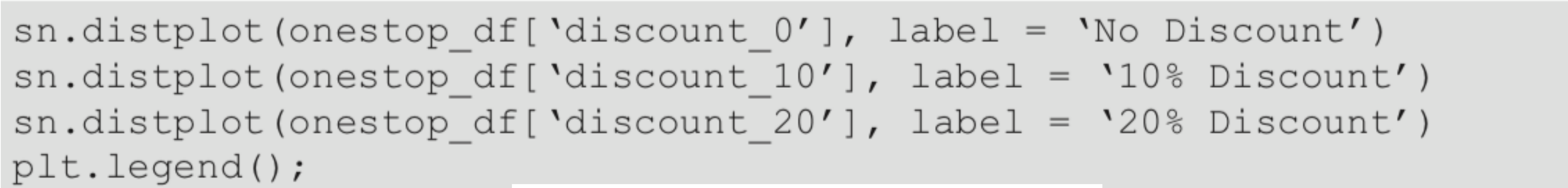


	discount_0	discount_10	discount_20
0	39	34	42
1	32	41	43
2	25	45	44
3	25	39	46
4	37	38	41

ANALYSIS OF VARIANCE (ANOVA) (Cntd.)

Solution:

Distribution plot of the groups



```
sn.distplot(onestop_df['discount_0'], label = 'No Discount')
sn.distplot(onestop_df['discount_10'], label = '10% Discount')
sn.distplot(onestop_df['discount_20'], label = '20% Discount')
plt.legend();
```

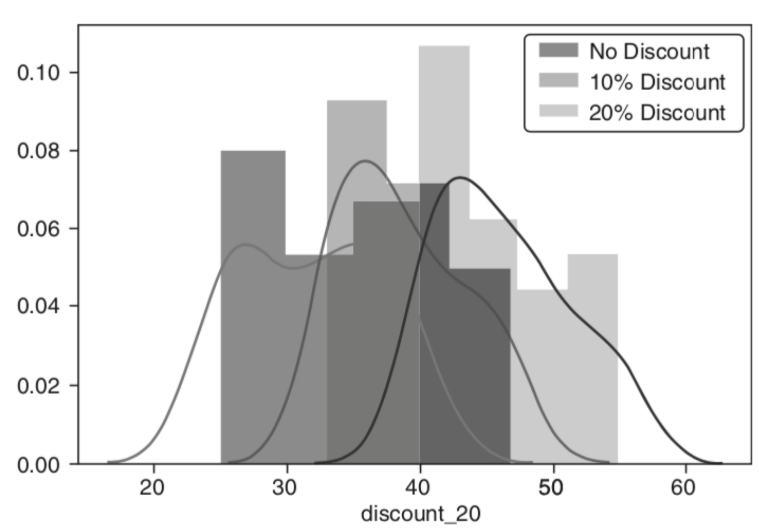
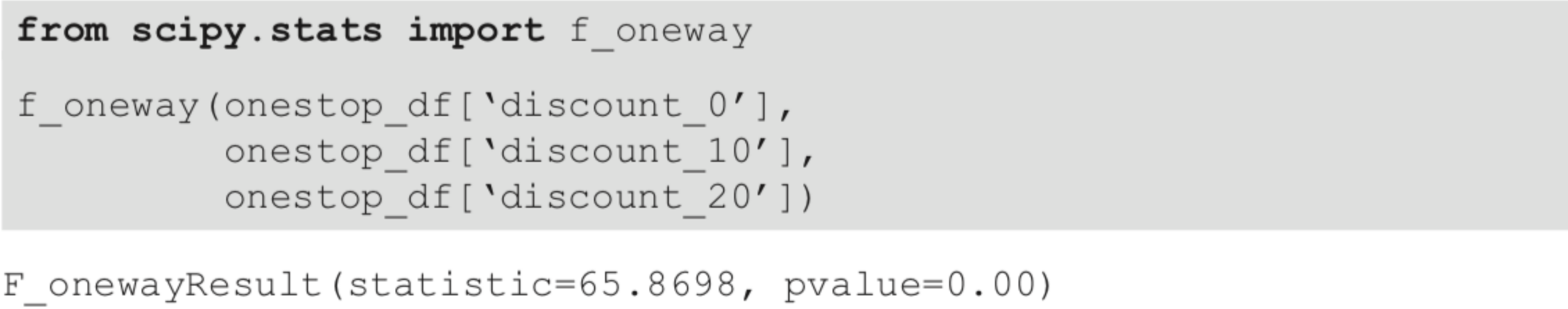


FIGURE 3.11 Comparing distributions of sales for different discount levels.

ANALYSIS OF VARIANCE (ANOVA) (Cntd.)

Solution:

- Conducting the test



```
from scipy.stats import f_oneway

f_oneway(onestop_df['discount_0'],
         onestop_df['discount_10'],
         onestop_df['discount_20'])
```

```
F_onewayResult(statistic=65.8698, pvalue=0.00)
```

- P-value is less than 0.05 value.
- We reject the null hypothesis.
- i.e., the mean sales quantity values under different discounts are different.

Thank You!