Hypothesis Testing:

You are a data analyst for a basketball team and have access to a large set of historical data that you can use to analyze performance patterns. The coach of the team and your management have requested that you perform several hypothesis tests to statistically validate claims about your team's performance.

Variable	What does it represent?
pts	Points scored by the team in a game
elo_n	A measure of relative skill level of the team in the league
year_id	Year when the team played the games
fran_id	Name of the NBA team

The ELO rating, represented by the variable `elo_n`, is used as a measure of the relative skill of a team. This measure is inferred based on the final score of a game, the game location, and the outcome of the game relative to the probability of that outcome. The higher the number, the higher the relative skill of a team.

```
In addition to studying data on your own team, your management has also assigned you a second team so that you can compare its performance with your own team's.
```

Team	What does it represent
Your Team	This is the team that has hired you as an analyst. This is the team that you will pick below. See Step 2.
Assigned Team	This is the team that the management has assigned to you to compare against your team. See Step 1.

Step 1: Data Preparation & the Assigned Team

This step uploads the data set from a CSV file. It also selects the Assigned Team for this analysis.

1. The **Assigned Team** is **Chicago Bulls** from the years **1996 - 1998**

In [ ]:
```
import numpy as np
import pandas as pd
import scipy.stats as st
import matplotlib.pyplot as plt
from IPython.display import display, HTML

nba_orig_df = pd.read_csv('nbaallelo.csv')
nba_orig_df = nba_orig_df[(nba_orig_df['lg_id']=='NBA') & (nba_orig_df['is_
columns_to_keep = ['game_id','year_id','fran_id','pts','opp_pts','elo_n','
nba_orig_df = nba_orig_df[columns_to_keep]

# The dataframe for the assigned team is called assigned_team_df.
# The assigned team is the Bulls from 1996-1998.
assigned_years_league_df = nba_orig_df[(nba_orig_df['year_id'].between(1996
assigned_team_df = assigned_years_league_df[(assigned_years_league_df['fran
assigned_team_df = assigned_team_df.reset_index(drop=True)

display(HTML(assigned_team_df.head().to_html()))
print("printed only the first five observations...")
print("Number of rows in the dataset =", len(assigned_team_df))
```

	game_id	year_id	fran_id	pts	opp_pts	elo_n	opp_elo_n	game_l
0	199511030CHI	1996	Bulls	105	91	1598.2924	1531.7449	
1	199511040CHI	1996	Bulls	107	85	1604.3940	1458.6415	
2	199511070CHI	1996	Bulls	117	108	1605.7983	1310.9349	
3	199511090CLE	1996	Bulls	106	88	1618.8701	1452.8268	
4	199511110CHI	1996	Bulls	110	106	1621.1591	1490.2861	

```
printed only the first five observations...
Number of rows in the dataset = 246
```

Step 2: Pick Your Team

In this step, you will pick your team. The range of years that you will study for your team is **2013-2015**.

In [ ]:
```
# Range of years: 2013-2015
# Note: The line below selects all teams within the three-year period 2013
# This is the dataframe for the other team; The Bulls.
your_years_leagues_df = nba_orig_df[(nba_orig_df['year_id'].between(2013, 2

# The dataframe for your team is called your_team_df.
# The your team is the Lakers from 2013-2015. This is to print only the fi
```

```
your_team_df = your_years_leagues_df[(your_years_leagues_df['fran_id']=='La
your_team_df = your_team_df.reset_index(drop=True)

display(HTML(your_team_df.head().to_html()))
print("printed only the first five observations...")
print("Number of rows in the dataset =", len(your_team_df))
```

	game_id	year_id	fran_id	pts	opp_pts	elo_n	opp_elo_n	game_l
0	201210300LAL	2013	Lakers	91	99	1541.7585	1533.9297	
1	201210310POR	2013	Lakers	106	116	1531.7184	1460.7015	
2	201211020LAL	2013	Lakers	95	105	1518.7981	1580.8679	
3	201211040LAL	2013	Lakers	108	79	1527.5927	1409.0566	
4	201211070UTA	2013	Lakers	86	95	1521.1603	1535.9674	

```
printed only the first five observations...
Number of rows in the dataset = 246
```

Step 3: Hypothesis Test for the Population Mean (I)

A relative skill level of 1340 represents a critically low skill level in the league. The management of your team has hypothesized that the average relative skill level of your team in the years 2013 - 2015 is greater than 1340. Test this claim using a 5% level of significance. Also, assume that the population standard deviation for relative skill level is unknown.

In [ ]:
```
import scipy.stats as st

# Mean relative skill level of your team
mean_elo_your_team = your_team_df['elo_n'].mean()
print("Mean Relative Skill of your team in the years 2013 to 2015 =", round

# Hypothesis Test
# Null Hypothesis: Mean relative skill level of your team in the years 201
test_statistic, p_value = st.ttest_1samp(your_team_df['elo_n'], 1340)

print("Hypothesis Test for the Population Mean")
print("Test Statistic =", round(test_statistic,2))
print("P-value =", round(p_value,4))
```

```
Mean Relative Skill of your team in the years 2013 to 2015 = 1440.49
Hypothesis Test for the Population Mean
Test Statistic = 19.8
P-value = 0.0
```

Step 4: Hypothesis Test for the Population Mean (II)

A team averaging 106 points is likely to do very well during the regular season. The coach of your team has hypothesized that your team scored at an average of less than 106 points in the years 2013-2015. Test this claim at a 1% level of significance. For this test, assume that the population standard deviation for relative skill level is unknown.

In [ ]:
```
# Calculate and print the mean points scored by the Lakers during the years
mean_pts_your_team = your_team_df['pts'].mean()
print("Mean points scored by your team in the years 2013 to 2015 =", round(

# Identify the mean score under the null hypothesis
null_hypothesis_value = 106

# Carry out the hypothesis test
test_statistic, p_value = st.ttest_1samp(your_team_df['pts'], null_hypothes

# Print the test statistic and P-value
print("Hypothesis Test for the Population Mean")
print("Test Statistic =", round(test_statistic,2))
print("P-value =", round(p_value,4))
```

```
Mean points scored by your team in the years 2013 to 2015 = 101.2
Hypothesis Test for the Population Mean
Test Statistic = -6.91
P-value = 0.0
```

Step 5: Hypothesis Test for the Population Proportion

Suppose the management claims that the proportion of games that your team wins when scoring 102 or more points is 0.90. Test this claim using a 5% level of significance.

In [ ]:
```
from statsmodels.stats.proportion import proportions_ztest

your_team_gt_102_df = your_team_df[(your_team_df['pts'] > 102)]

# Number of games won when your team scores over 102 points
counts = (your_team_gt_102_df['game_result'] == 'W').sum()

# Total number of games when your team scores over 102 points
nobs = len(your_team_gt_102_df['game_result'])

p = counts*1.0/nobs
print("Proportion of games won by your team when scoring more than 102 poi
```

```
# Conduct a hypothesis test using the test statistic and P-value to calcula
# Assuming that the population standard deviation is unknown;
# The null hypothesis is that the proportion of games that your team wins u
test_statistic, p_value = proportions_ztest(counts, nobs, null_hypothesis_v

print("Hypothesis Test for the Population Proportion")
print("Test Statistic =", round(test_statistic,2))
print("P-value =", round(p_value,4))
```

```
Proportion of games won by your team when scoring more than 102 points in t
he years 2013 to 2015 = 0.5268
Hypothesis Test for the Population Proportion
Test Statistic = -2235.66
P-value = 0.0
```

Step 6: Hypothesis Test for the Difference Between Two Population Means

The management of your team wants to compare the team with the assigned team (the Bulls in 1996-1998). They claim that the skill level of your team in 2013-2015 is the same as the skill level of the Bulls in 1996 to 1998. In other words, the mean relative skill level of your team in 2013 to 2015 is the same as the mean relative skill level of the Bulls in 1996-1998. Test this claim using a 1% level of significance. Assume that the population standard deviation is unknown.

In [ ]:
```
import scipy.stats as st

mean_elo_n_project_team = assigned_team_df['elo_n'].mean()
print("Mean Relative Skill of the assigned team in the years 1996 to 1998 =

mean_elo_n_your_team = your_team_df['elo_n'].mean()
print("Mean Relative Skill of your team in the years 2013 to 2015  =", roun


# Conducting a Hypothesis Test for the difference between two population me
test_statistic, p_value = st.ttest_ind(assigned_team_df['elo_n'], your_tean

print("Hypothesis Test for the Difference Between Two Population Means")
print("Test Statistic =", round(test_statistic,2))
print("P-value =", round(p_value,4))
```

```
Mean Relative Skill of the assigned team in the years 1996 to 1998 = 1739.8
Mean Relative Skill of your team in the years 2013 to 2015  = 1440.49
Hypothesis Test for the Difference Between Two Population Means
Test Statistic = 49.51
P-value = 0.0
```

End of Hypothesis Testing