Individual Player Analysis

Now that our data is properly configured, we can begin with our first stage of analysis. We will be investigating if, on the individual player level, the result of the first free throw affects the probability of making the second free throw.

We will do this by performing hypothesis tests with the null hypothesis H_0: Free throw 2 of 2 is independent of free throw 1 of 2. In other words, knowing the result of the first free throw does not affect the probability of making the second free throw.

[33]:
```
import pandas as pd
import scipy.stats as st
import numpy as np

#Create a DataFrame by reading CSV
df = pd.read_csv(r'free_throw_counts.csv')

df.head(5)
```

[33]:
```
            player  missed_1st_made_2nd  missed_first  made_1st_made_2nd  \
0      A.J. Price                   21            28                 62
1    Aaron Brooks                   69            83                295
2    Aaron Gordon                   27            38                 53
3      Aaron Gray                   32            65                 51
4  Aaron Harrison                    2             3                  0

   made_first
0          77
1         345
2          72
3          87
4           3
```

Because we will be performing multiple hypothesis tests for different players, we will need to adjust our significance level to account for the increased likelihood of making a Type I error (false positive) across multiple tests.

When we adjust our significance level, it will reduce the power of our individual hypothesis tests. So, we will only perform tests for the 50 players with the largest total number of free throws.

[34]:
```
#Create a new column for the sum of 'missed_first' and 'made_first'
df['total'] = df['missed_first'] + df['made_first']

#Sort the DataFrame by the 'total_shots' column in descending order
df = df.sort_values(by='total', ascending=False)

#Remove the 'total' column
df.drop('total', axis='columns', inplace=True)
```

```
#Only keep the 50 players with the largest total number of free throws
df=df.head(50)

df.head(5)
```

[34]:

```
            player  missed_1st_made_2nd  missed_first  made_1st_made_2nd  \
293  Dwight Howard                  853          1589               1134
607   LeBron James                  730           951               1871
563   Kevin Durant                  284           322               1854
295    Dwyane Wade                  472           622               1424
576    Kobe Bryant                  333           382               1607

     made_first
293        1887
607        2389
563        2086
295        1784
576        1872
```

We will assume that pairs of free throws are independent of each other and that our probabilities of success within each group (made/missed first free throw) remain constant for all trials.

Let the random variable Y_1 represent the number of successful 2nd free throw attempts in a sample of n_1 free throw pairs, where the 1st free throw was unsuccessful, drawn from a large population.

Based on our assumptions, Y_1 can be closely modeled by a binomial distribution $BIN(\theta_1)$, where θ_1 represents the probability of making the 2nd free throw after missing the 1st free throw.

Our maximum likelihood estimate for θ_1 is $\hat{\theta}_1 = \frac{y_1}{n_1}$.

As an example, our observed value for LeBron James is $y_1 = 730$ and $n_1 = 951$, giving us $\hat{\theta}_1 = \frac{730}{951} \approx 0.7676$.

Defining Y_2 as the number of successful 2nd free throw attempts in a sample of n_2 free throw pairs where the 1st free throw was successful and applying the same process, we get $\hat{\theta}_2 = \frac{1871}{2389} \approx 0.7832$, where θ_2 is defined as the probability of making the 2nd fee throw after making the 1st free throw.

We would like to see if the difference between θ_1 and θ_2 is statistically significant at the 0.05 significance level.

Our null hypothesis is $H_0 : \theta_1 = \theta_2$.

From the National Institute of Standards and Technology (www.itl.nist.gov/div898/handbook/prc/section3/prc33.htm), we have that $\frac{\tilde{\theta}_1 - \tilde{\theta}_2}{\sqrt{\tilde{\theta}(1-\tilde{\theta})(\frac{1}{n_1}+\frac{1}{n_2})}} \sim G(0,1)$, where $\tilde{\theta} = \frac{n_1\tilde{\theta}_1 + n_2\tilde{\theta}_2}{n_1+n_2}$

Our test statistic is $D = |Z|$, and our observed value of D is $d \approx 0.9777$.

Our approximate p-value based on the Gaussian approximation is

$p - value = P(D \geq d; H_0)$

$= P(\frac{|\tilde{\theta}_1 - \tilde{\theta}_2|}{\sqrt{\tilde{\theta}(1-\tilde{\theta})(\frac{1}{n_1}+\frac{1}{n_2})}} \geq 0.9777)$
$\approx P(|Z| \geq 0.9777)$
$= 2[1 - P(Z \leq 0.9777)]$
$\approx 2(1 - 0.8359)$
≈ 0.3282

Since our p-value is greater than 0.05, we conclude that there is insufficient evidence to reject the null hypothesis for LeBron James.

We can repeat this process for each of our 50 players using python.

[35]:
```
df['theta_hat_1'] = df['missed_1st_made_2nd'] / df['missed_first']

df['theta_hat_2'] = df['made_1st_made_2nd'] / df['made_first']

df['theta_tilde'] = ((df['missed_first'] * df['theta_hat_1'] +
                      df['made_first'] * df['theta_hat_2']) /
                      (df['missed_first'] + df['made_first']))

df['d'] = np.abs((df['theta_hat_1'] - df['theta_hat_2']) /
                 np.sqrt(df['theta_tilde'] * (1-df['theta_tilde']) *
                         (1/df['missed_first'] + 1/df['made_first'])))

df['p-val'] = 2 * (1 - st.norm.cdf(df['d']))

df.head(5)
```

[35]:
```
             player  missed_1st_made_2nd  missed_first  made_1st_made_2nd  \
293  Dwight Howard                  853          1589               1134
607   LeBron James                  730           951               1871
563   Kevin Durant                  284           322               1854
295    Dwyane Wade                  472           622               1424
576    Kobe Bryant                  333           382               1607

     made_first  theta_hat_1  theta_hat_2  theta_tilde         d     p-val
293        1887     0.536816     0.600954     0.571634  3.806789  0.000141
607        2389     0.767613     0.783173     0.778743  0.977652  0.328247
563        2086     0.881988     0.888782     0.887874  0.359669  0.719095
295        1784     0.758842     0.798206     0.788030  2.068395  0.038603
576        1872     0.871728     0.858440     0.860692  0.683505  0.494288
```

We can already see a few players that seem to provide sufficient evidence to reject the null hypothesis at the 0.05 level. However, we still need to make some adjustments to account for multiple hypothesis tests.

To do this, we will use the Benjamini-Hochberge Procedure (https://www.statisticshowto.com/benjamini-hochberg-procedure/).

[36]:
```python
#Sort the rows from smallest to largest p-value
df = df.sort_values(by='p-val', ascending=True)

#Reset the dataframe indices
df = df.reset_index()
df.drop('index', axis='columns', inplace=True)

#Add  the critical value as a column
df['critical val'] = ((df.index + 1) / 50) * 0.05

df.head(5)
```

[36]:
```
              player  missed_1st_made_2nd  missed_first  made_1st_made_2nd  \
0      Dwight Howard                  853          1589               1134
1         Josh Smith                  380           628                629
2     Corey Maggette                  192           241                881
3     Andre Iguodala                  345           496                752
4  LaMarcus Aldridge                  220           293                982

   made_first  theta_hat_1  theta_hat_2  theta_tilde         d     p-val  \
0        1887     0.536816     0.600954     0.571634  3.806789  0.000141
1         903     0.605096     0.696567     0.659046  3.713779  0.000204
2        1016     0.796680     0.867126     0.853620  2.781425  0.005412
3         990     0.695565     0.759596     0.738223  2.647787  0.008102
4        1201     0.750853     0.817652     0.804552  2.585265  0.009730

   critical val
0         0.001
1         0.002
2         0.003
3         0.004
4         0.005
```

[37]:
```python
#Find the largest p-value less than or equal to their critical value
pval = (df[df['p-val'] <= df['critical val']]).max()['p-val']

#Players with p-values less than or equal to this value are significant
df_stat_sig = df[df['p-val'] <= pval]

print(df_stat_sig)
```

```
          player  missed_1st_made_2nd  missed_first  made_1st_made_2nd  \
0  Dwight Howard                  853          1589               1134
1     Josh Smith                  380           628                629

   made_first  theta_hat_1  theta_hat_2  theta_tilde         d     p-val  \
0        1887     0.536816     0.600954     0.571634  3.806789  0.000141
1         903     0.605096     0.696567     0.659046  3.713779  0.000204
```

```
   critical val
0         0.001
1         0.002
```

Therefore, of our 50 players, only Dwight Howard and Josh Smith have statistically significant differences between the probability of them making the second free throw after making and after missing the first free throw.

Both are more likely to make the second free throw if they make the first free throw.

Dwight Howard is approximately 6.4% more likely to make the second free throw if he makes the first.
Josh Smith is approximately 9.1% more likely to make the second free throw if he makes the first.

So, not only is the difference statistically significant for these players, the difference is practically significant.