SZEGEDI TUDOMÁNYEGYETEM

Természettudományi és Informatikai Kar

Bolyai Intézet

Matematika BSc

Futball Eredmények Előrejelzésének Statisztikai Vizsgálata

Tamás Tibor

2023.05.18

Tartalomjegyzék

Date	HomeTeam	AwayTeam	FTHG	FTAG	HxG	AxG
23Aug14 – 23May15	20 unique values	20 unique values	0 – 9	0 – 8	0.08 – 6.61	0.04 – 6.19
2014-08-23	Almeria	Espanyol	1	1	0.98	0.4
2014-08-23	Granada	Deportivo La Coruna	2	1	0.55	0.38
2014-08-23	Malaga	Athletic Club	1	0	1.32	1.14
2014-08-23	Sevilla	Valencia	1	1	1.17	1.75
2014-08-24	Barcelona	Elche	3	0	1.54	0.11
2014-08-24	Celta Vigo	Getafe	3	1	1.92	0.67
2014-08-24	Eibar	Real Sociedad	1	0	0.47	0.98
2014-08-24	Levante	Villarreal	0	2	0.35	3.01
2014-08-25	Real Madrid	Cordoba	2	0	0.61	0.38
2014-08-25	Rayo Vallecano	Atletico Madrid	0	0	0.36	0.9
2014-08-29	Getafe	Almeria	1	0	1.32	1.31
2014-08-29	Valencia	Malaga	3	0	1.97	1.13
2014-08-30	Athletic Club	Levante	3	0	2.26	0.33

Figure 1: Nyers adat a használt statisztikákkal

1 Bevezetés

A projekt során a Kaggle platformról származó adatokon fogok különböző statisztikákat számítani, összehasonlítani különböző adathalmazokat illetve egy később felvetendő kérdésre szeretnék választ kapni amihez hipotézis vizsgálatot fogok segítségül hívni.

2 Adatok bemutatása

Az adat a Kaggle platformról származik és az öt legnagyobb európai futball ligáról tartalmaz statisztikákat 2014 és 2022 között. A teljes adat halmaz több mint 10000 meccs eredményeit és statisztikáit tartalmazza (összesen 64 különböző statisztika meccsenként). A projektben csak a spanyol első osztályú mérkőzéseket fogom vizsgálni és csak 4 különböző statisztikát fogok használni. A nyers adathalmaz a 1 ábrán látható.

A használt statisztikák a következők:

1. FTHG - a hazai csapat góljainak száma
2. FTAG - a vendég csapat góljainak száma
3. HxG - a hazai csapat góljainak előrejelzése
4. AxG - a vendég csapat góljainak előrejelzése

A projekt célja, hogy megállapítsuk helyesek-e a fenti becslések a gólokról.

	Mean	Median	Skewness	Kurtosis
FTHG	1.50559	1	1.09735	1.85072
FTAG	1.13355	1	1.27784	2.44097
HxG	1.50824	1.34	1.21173	2.36929
AxG	1.13052	0.96	1.34756	2.67939

2.1 Adatok feldolgozása

Az adatok tisztítva lettek feltöltve így további adat előfeldolgozásra nem volt szükség a projekthez. A különböző évek adatai külön CSV fájlokban voltak tárolva ezeket kellett egyesítenem.

2.2 Leíró statisztikák

Az adatoknak négy leíró statisztikáját vizsgáltam amelyek a következők:

1. Átlag
2. Medián
3. Skewness (ferdeség)
4. Kurtosis (csúcsosság)

A kapott értékek a 2.2 táblázatban láthatóak. A kapott eredményekből a következő tapasztalatokat vonhatjuk le:

1. A hazai gólokat felül míg a vendég gólokat alul becsülték
2. Mivel egyik ferdeség sem egyenlő nullával ezért már itt feltehetjük, hogy nem normális eloszlásról beszélünk.
3. Mivel a csúcsosság mindenhol kisebb mint 3 ezért ez megerősíti az előző feltevésünket mi szerint nem normális eloszlásúak az adataink.

2.3 Hisztogram

Miután elkészítettük az adatok hisztogramját 2 láthatjuk, hogy valóban nem normális eloszlásról beszélünk.

2.4 Boxplot

Miután elkészítettük a box plotokat 3 ezek is megerősítették a feltevésünket.

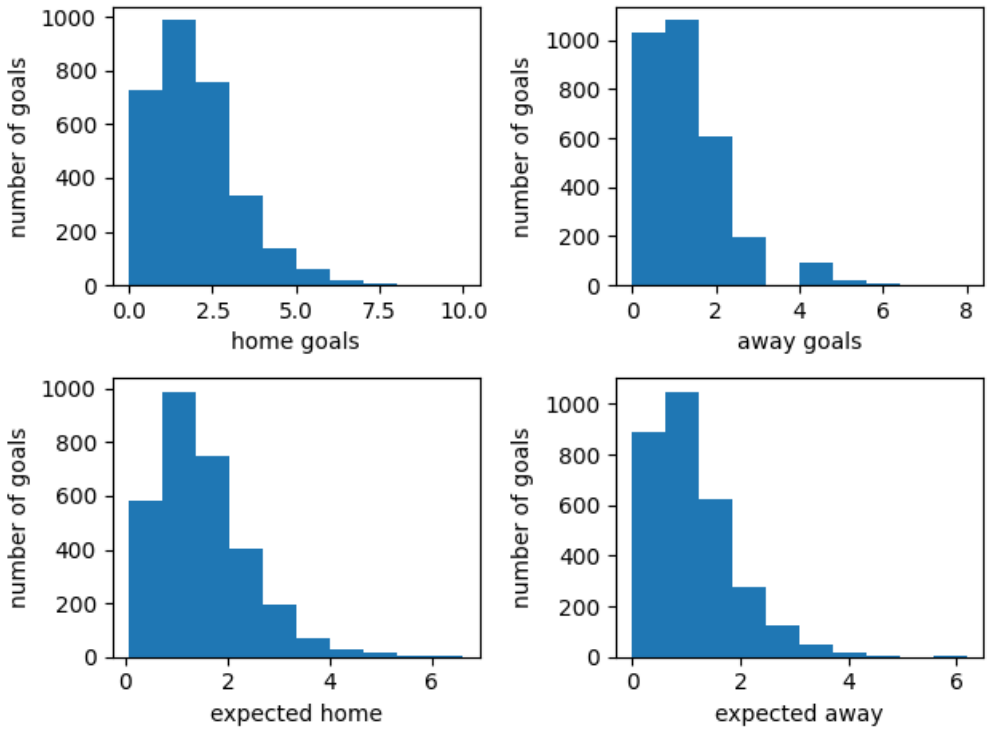


Figure 2: A változók hisztogramjai

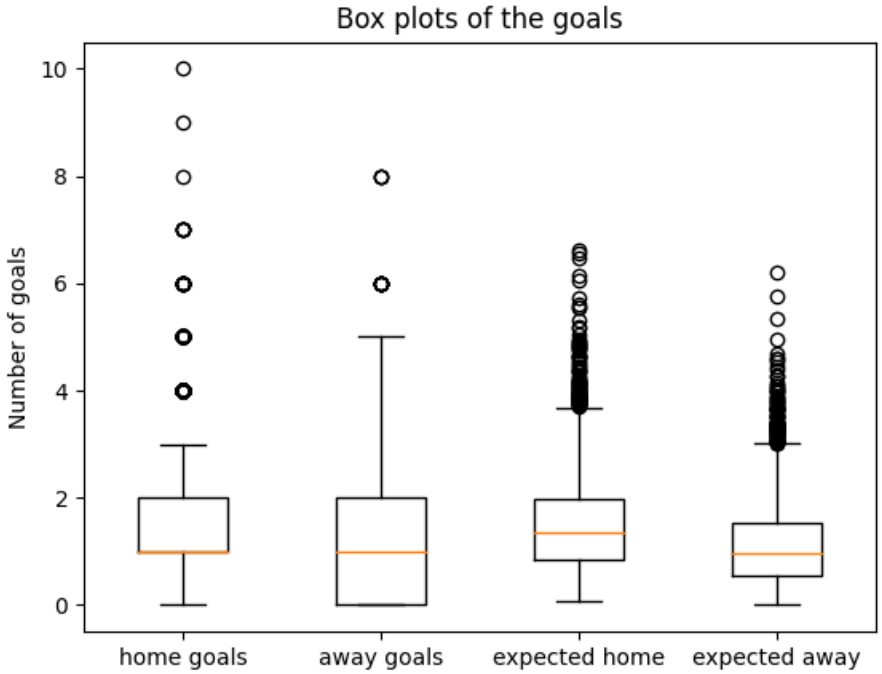


Figure 3: A változók box plotjai

2.5 Q-Q plot

Hogy biztosak legyünk benne, hogy nem normális eloszlással van dolgunk végül egy elkészültek a változók Q-Q plotjai is amikről ez egyértelműen látszódik.2.5

3 Korreláció vizsgálat

Az adatokat párokba állítottuk és ezek alapján vizsgáltuk a korrelációt köztük:

1. Hazai gólok száma - becsült hazai gólok száma.
2. Vendég gólok száma - becsült vendég gólok száma.

A korreláció vizsgálatához Pearson illetve Spearman-féle korrelációs tesztet végeztem

3.1 Pearson-féle korrelációs teszt

A nullhipotézis, hogy páronként a változók függetlenek, a teszt után a következő eredményeket kaptam:

1. 1. teszt : $\boldsymbol{p = 1.2e^{-12}}$ és a teszt statisztika = 0.6481
2. 2. teszt : $\boldsymbol{p = 1.32e^{-12}}$ és a teszt statisztika = 0.6444

Ezekből megállapíthatjuk, hogy páronként a két változó egymással pozitív korrelációban áll, tehát a null hipotézist elvetjük.

3.2 Spearman-féle korrelációs teszt

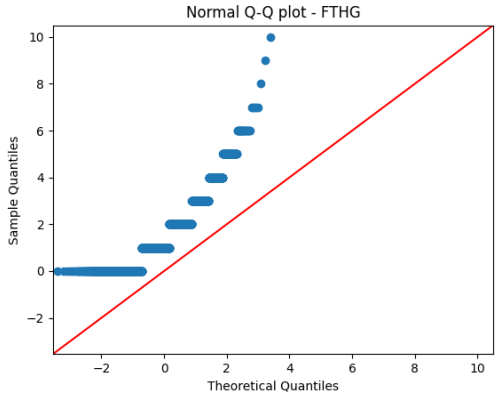
A Pearson-féle teszthez hasonlóan páronként elvégeztük a Spearman-féle rangkorrelációs tesztet amelyre a következő eredmény kaptuk:

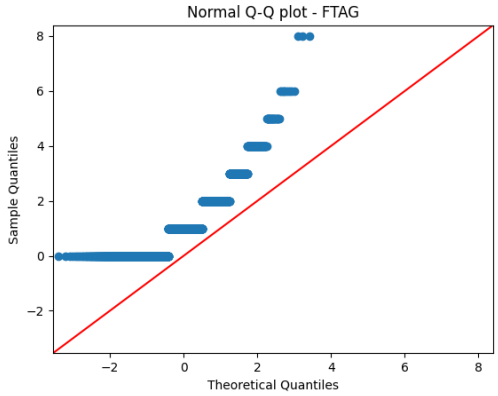
1. 1. teszt : $\boldsymbol{p = 1.126e^{-310}}$ és a teszt statisztika = 0.6110
2. 2. teszt : $\boldsymbol{p = 1.498e^{-294}}$ és a teszt statisztika = 0.5982

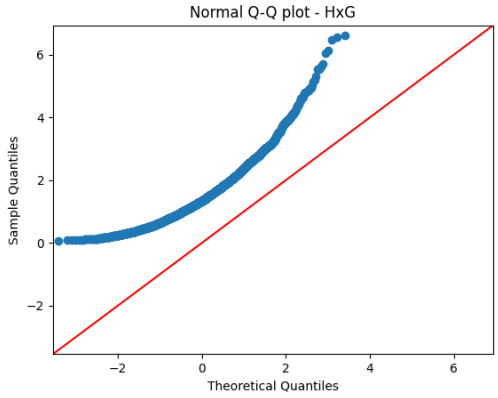
Ebben a tesztben is elutasítjuk a null hipotézist mi szerint a változók nem korreláltak.

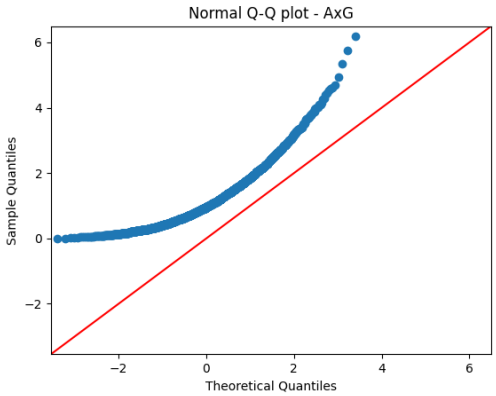
3.3 A korrelációra vonatkozó megállapítások

A fent elvégzett tesztek alapján megállapíthatjuk, hogy a páronként vizsgált változók egymással erősen, pozitívan korreláltak.


Normal Q-Q plot - FTHG
Sample Quantiles
Theoretical Quantiles


Normal Q-Q plot - FTAG
Sample Quantiles
Theoretical Quantiles


Normal Q-Q plot - HxG
Sample Quantiles
Theoretical Quantiles


Normal Q-Q plot - AxG
Sample Quantiles
Theoretical Quantiles

4 Hipotézis vizsgálat

Azt a nullhipotézist állítjuk fel, hogy nincs számottevő eltérés a valós gólok és a becsült gólok száma között. A hipotézis vizsgálathoz is, ahogy a korreláció vizsgálathoz párokba rendeztük a változókat:

1. Hazai gólok száma - becsült hazai gólok száma.
2. Vendég gólok száma - becsült vendég gólok száma.

4.1 Két mintás t-próba

A hipotézis vizsgálathoz páros t-próbát végeztünk. Mivel nagy elemszámú a mintánk ezért annak ellenére, hogy nem normális eloszlásúak a változóink erre lehetőségünk van. Elvégezve a próbát a következő eredményeket kaptuk:

1. 1. teszt : $\boldsymbol{p = 0.9270}$ és a teszt statisztika = -0.0915
2. 2. teszt : $\boldsymbol{p = 0.9041}$ és a teszt statisztika = 0.1204

Ezek alapján a null hipotézist nem tudjuk elvetni, tehát nincs számottevő eltérés a változók várható értékében.

5 Konklúzió

Az elvégzett számítások és a hipotézis vizsgálat alapján arra következtetésre jutottunk, hogy a becslések nagy valószínűséggel helyesek. Ám míg a gólok valós száma csak pozitív egész számokat vehet fel addig a becslések mind pozitív valós számok ezért az egy további projekt célja lehetne, hogy mi történik ezekkel a számításokkal ha mondjuk minden becsült értéknek az egész részét vagy a kerekítését vesszük.

6 Számítási módszerek

A projektben elvégzett számításokhoz a Python scipy.stats illetve statmodels csomagjait használtam plotoláshoz pedig a matplotlib.pyplot csomagot. A teljes kód elérhető a következő GitHub repositoryban.