Анализа података ЕнБиЕј (енг. *NBA*) лиге у периоду 1996-2020.

Семинарски рад у оквиру курса
Увод у теорију узорака
Математички факултет

Никола Јанковић

15. октобар 2020.

Сажетак

Рад представља симулацију истраживања над популацијом ЕнБиЕј играча у периоду 1996-2019, у нади да ће бити откривене неке значајне популацијске вредности као и покушај да се докажу неке потврде неке хипотезе, као што су да ли је могуће имати квалитетно академско образовање и професионалну кошаркашку каријеру, да ли су амерички играчи повлашћенији на пут до лиге и сличне.

Садржај

1 Увод

Спорту као таквом увелико примарна улога није оно што је на почетку била, а то је забава великог броја људи. Већ је се претворило у озбиљну индустрију, што само по себи значи велику количину новца као крајњи производ. Како количина новца која постоји у професионалном спорту сваким даном расте тако је и потреба за његовим унапређењем порасла. Спортске организације улажу огромне своте новца у прикупљање и анализу велике количине података о самом процесу игре, како би пронашли потенцијалне недостатке и унапредили недостатке који на основу простог гледања није могуће уочити, а други разлог је чињеница да крајњи корисници тј. љубитељи спорта воле да прате такве информације скоро исто као и саму игру.

У раду ће бити приказане неке методе оцењивања неких једноставних статистчких параметара, као и неких мало комплекснијих над скупом података о играчима ЕнБиЕј лиге током периода од 20-година. Скуп података није ни близу детаљан као неки други скупови на ову тему, који или нису доступни јавно или захтевају много више времена на припреми података. Што свакако није циљ курса у склопу ког је писан овај рад.

Рад је подељен у три целине. У првом делу ће бити обрађене опште познате и популарне статистике кошаркашке игре, као што су просечан број поена, асистенција и ухваћених лопти. Биће коришћена два приступа при узорковању. Други део чини оцена неких напреднијих статистика, где ће се као методе узорковања користити систематски узорак, регресионо оцењивање и количничко оцењивање. И на самом крају, оценићемо пропорције неких категоричких обележја помоћу узроковања стратрификоване популације и простог случајног узорковања са понављањем.

Мотив за избор овог скупа података дат је на почетку ове секције. Као референтне изворе литературе наводимо [1, 2]

2 Опис базе података

Популација над којом ће бити вршено истраживање представљају играчи који су играли у америчко-канадској националној кошаркашкој лиги у периоду од 1996. године до 2019. Пошто је база података потпуна тиме ће цела популација представљати у исто време и оквир за одабир узорка. Док јединицу узорковања представља кошаркаш и његове значајне статистике током одређене сезоне. Дакле играч А током сезоне 2003/2004 и сезоне 2004/2005 представља посебан ентитет.

Обим популације једнак је 11145 и сваки појединачни ентитет је сачињен од 22 обележја. За сад ћемо само набројати та обележја: име, назив тима, број година које је играч имао током те сезоне, висина, тежина, похађан колец, држављанство, година у којој је позван да дође у лигу (енг. *draft year*), рунда у којој је изабран (енг. *draft round*), редни број при избору (енг. *draft number*), број одиграних утакмица те сезоне (енг. *Game Played*), просечан број поена, асистенција и ухваћених лопти током сезоне, сезона и још неколико специфичних статистика: нет_рејтинг, ореб_пцт, дреб_пцт, тс_пцт, аст_пцт. За које аутор сматра да је много сврсисходније их детаљније појаснити у делу у ком буду коришћени.

Иако је база садржи читаву популацију популацијске статистике неће бити изнете јер је циљ рада симулација реалног истраживања у

ком не испитујемо све јединке популације. Сем у једном случају који ће јасно бити наглашен касније.

Преглед базе је дат на адреси

3 Поређење основних нумеричких статистика

Често, у круговима љубитеља ЕнБиЕј лиге, влада мишљење да се последњих неколико сезона играчи мање труде при обављању својих задатака у фази одбране како би се одморили и смањили потенцијалну шансу да агресивним приступом доведу себе у проблем са личним прешкама и буде им редукована минутажа, а самим тим и број постигнутих поена. Упоредићемо просечан број поена играча током сезоне 2019/2020 и свих осталих сезона, да бисмо видели да је разлика примента. Очекивано је да просечан број поена буде већи данас него претходних 13 година, асистенције на сличном нивоу, док се за ухбаћене лопте очекује да буду у порасту услед брже игре и већег броја покушаја током утакмице, а самим тим и већег броја промашаја.

3.1 Прост случајан узорак без понављања

Непристрасна оцена просечног броја поена играча, асистенција и скокова респективно током периода 1996-2020 методом простог случајног узорка без понављања је дато формулом:

$$\overline{x} = \frac{1}{n} \sum_{k \in S} x_k$$

где S представља узорак из оквира играча, а x_k представља вредност горе поменутих обележја на појединачној јединици узорковања. Док је оцена дисперзије дате оцене дата формулом:

$$\widehat{D(\overline{x})} = \frac{s_n^2}{n}\left(1 - \frac{n}{N}\right)$$

где s_n^2 представља узорачку дисперзију, n обим узорка, а N обим популације. https://www.overleaf.com/project/5f63d1f1e7edbd0001a157a2

3.2 Кластер узорак код ког се примарне јединице бирају као прост случајан узорак

Непристрасна оцена просечног броја поена играча, асистенција и скокова респективно током периода 1996-2020 методом кластер узорковања код ког се примарне јединице бирају као прост случајан узорак задата је формулом:

$$\overline{x} = \frac{\frac{N}{n} \sum\limits_{i \in S} t_i}{\sum\limits_{j=1}^{N} Mi}$$

Док је оцена дисперзије дате оцене дата формулом:

$$\widehat{D(\overline{x})} = N^2\left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{1}{n-1}\sum_{i \in S}\left(t_i - \frac{1}{n}\sum_{j \in S} t_j\right)^2$$

где N представља број примарних јединица у популацији. У овом примеру једну примарну јединицу чине играчи који су играли током конкретне сезоне. Док је n број примарних јединица изабраних у узорак.

3.3 Резултати узорковања

Уз горе наведену популацију при истраживању је коришћен програмски језик Ар (енг. *R*). Фрагмент Ар кода који рачуна тражене статике играча током сезоне 2019-2020, оцене тих статистика методом простог случајног узорка без понављања и кластер узорка следи у наставку : [1]

```
est.d.srswor = function(sampled.data, n, N) {
    return(1/n * var(sampled.data) * (1 - n/N))
}
srswor.1920.vs.all.seasons.sample = function(n) {
    #
     ##############################################################

    idxs = sample(seq(nrow(nbaPlayers)), size = n, replace = FALSE
    )
    nba.sample = nbaPlayers[idxs, ]
    pts.sample = nba.sample$pts
    cat('Prosecan broj poena u sezoni 2019-20: ',
    mean(data.19.20$pts), '\n')
    cat('Procenjen prosecan broj poena u prethodnih 13 sezona: ',
    mean(pts.sample), '\n')
    cat('Ocena disperzije ocene je: ',
    est.d.srswor(sampled.data = pts.sample, n, nrow(nbaPlayers)),
    '\n')
    cat('################################################\n')
    ##############################################################
    ast.sample = nba.sample$ast
    cat('Prosecan broj asistencija u sezoni 2019-20: ',
    mean(data.19.20$ast), '\n')
    cat('Procenjen prosecan broj asistencija u prethodnih 13
    sezona: ',
    mean(ast.sample), '\n')
    cat('Ocena disperzije ocene je: ',
    est.d.srswor(sampled.data = ast.sample, n, nrow(nbaPlayers)),
    '\n')
    cat('################################################\n')
    ##############################################################
    reb.sample = nba.sample$reb
    cat('Prosecan broj skokova u sezoni 2019-20: ',
    mean(data.19.20$reb), '\n')
    cat('Procenjen prosecan broj skokova u prethodnih 13 sezona: '
    ,
    mean(reb.sample), '\n')
    cat('Ocena disperzije ocene je: ',
    est.d.srswor(sampled.data = reb.sample, n, nrow(nbaPlayers)),
    '\n')
}

filter.clusters = function(n = 1, N = 1) {
    idxs = sample(0:N-1, n, replace = F)
    idxs
    seasons = c()
    for (i in 1:n) {
        seasons = c(seasons, paste(c(1996 + idxs[i], substr((1996
    + idxs[i] + 1), 3, 4)), collapse = '-'))
    }
    cat(length(unique(seasons)))
    return(nbaPlayers %>% filter(season %in% seasons))
}

```

[1] У току рада коришћена је библиотека *DPLYR*

Табела 1: Резултати узорковања са ПСУБП и КУПСУ

	поени	асистенције	скокови
2019-20	8.626654	1.901556	3.601556
ПСУБП	8.527	1.758	3.684
$\widehat{D(\hat{\theta})}$	0.3525532	0.02490418	0.05607958
КУПСУ	7.638342	1.718499	3.384581
$\widehat{D(\hat{\theta})}$	0	0	-6.666229e-20

```
#Function which returns estimated value of mean and estimated
     variance of esitmated value of mean
est.cluster.srswor = function(clusters, N, n) {
    ti.pts = c()
    ti.ast = c()
    ti.reb = c()
    seasons = unique(clusters$season)
    for (s in seasons) {
        filtered.cluster = clusters %>% filter(season == s) %>%
     select(pts, ast, reb, season)
        cluster.sample = (filtered.cluster
             %>% summarise(pts = sum(pts),
                           ast = sum(ast),
                           reb = sum(reb)))

        ti.pts = c(ti.pts, cluster.sample$pts)
        ti.ast = c(ti.ast, cluster.sample$ast)
        ti.reb = c(ti.reb, cluster.sample$reb)
    }
    cat('################################################\n')
    cat('Klaster ocenjen prosecan broj poena igraca: ',
    N  / n * sum(ti.pts) / nrow(nbaPlayers), '\n')
    cat('Ocena disperzije klaster ocene poena: ', 1 / (nrow(
     nbaPlayers)^2) * N^2 * (1 - n/N) * 1/n * 1/(n-1) * sum(ti.pts
      - 1/n * sum(ti.pts)),  '\n')
    cat('Klaster ocenjen prosecan broj asistencija igraca: ',
    N  / n * sum(ti.ast) / nrow(nbaPlayers), '\n')
    cat('Ocena disperzije klaster ocene asistencija: ',
    1 / (nrow(nbaPlayers)^2) * N^2 * (1 - n/N) * 1/n * 1/(n-1) *
     sum(ti.ast - 1/n * sum(ti.ast)),  '\n')
    cat('Klaster ocenjen prosecan broj skokova igraca: ',
    N  / n * sum(ti.reb) / nrow(nbaPlayers), '\n')
    cat('Ocena disperzije klaster ocene skokova: ',
    1 / (nrow(nbaPlayers)^2) * N^2 * (1 - n/N) * 1/n * 1/(n-1) *
     sum(ti.reb - 1/n * sum(ti.reb)), '\n')
}
```

Резулатати добијени код ПСУБП[2] за $n = 100$ и КУПСУ [3] за $n = 4$ приказани су у табели 3.3. Као што видимо оцене дисперзија су боље у случају кластер узорка, па ако оцене добијене кластер узорком као релеватнију вредност увиђамо да су наше претпоставке биле добре. Поен више у просеку сваког играча на утакмици, знајући да се тим састоји од 12 играча значи да тимови на сваком мечу у просеку постижу 12 поена више, што узевши у обзир да тимови постижу од 80 до 120 поена на утакмици је заиста значајна разлика.

[2]Прост случајан узорак без понављања

[3]Кластер узорак где се примарне јединице бирају методом простог случајног узорка без понављања

3.4 Интервална оцена броја играча по тиму током периода 1996-2020

Како су многи тимови мењали град, тако у бази постоји неконзистентност по питању броја играча у тимовима. Неопходно је било прво кориговати тај проблем овим фрагментом кода.

```
team.num.of.players = nbaPlayers %>%
distinct(player_name, team_abbreviation) %>%
group_by(team_abbreviation) %>%
summarise(nrow = n())

n = sum(team.num.of.players[which(team.num.of.players$team_
    abbreviation %in% c('BKN', 'NJN')), ]$nrow)
team.num.of.players = rbind(team.num.of.players, data.frame(team_
    abbreviation = "BKN/NJN", nrow = n))
n = sum(team.num.of.players[which(team.num.of.players$team_
    abbreviation %in% c('VAN', 'MEM')), ]$nrow)
team.num.of.players = rbind(team.num.of.players, data.frame(team_
    abbreviation = "VAN/MEM", nrow = n))
n = sum(team.num.of.players[which(team.num.of.players$team_
    abbreviation %in% c('NOP', 'NOH', 'NOK')), ]$nrow)
team.num.of.players = rbind(team.num.of.players, data.frame(team_
    abbreviation = "NOP/NOH/NOK", nrow = n))
n = sum(team.num.of.players[which(team.num.of.players$team_
    abbreviation %in% c('CHA', 'CHH')), ]$nrow)
team.num.of.players = rbind(team.num.of.players, data.frame(team_
    abbreviation = "CHA/CHH", nrow = n))
n = sum(team.num.of.players[which(team.num.of.players$team_
    abbreviation %in% c('SEA', 'OKC')), ]$nrow)
team.num.of.players = rbind(team.num.of.players, data.frame(team_
    abbreviation = "SEA/OKC", nrow = n))
team.num.of.players =  team.num.of.players %>% filter(!(team_
    abbreviation %in% c('BKN', 'NJN', 'VAN', 'MEM', 'NOP', 'NOH',
     'NOK', 'CHA', 'CHH', 'SEA', 'OKC')))
```

Интервална оцена за популацијску средњу вредност (тимова постоји 30, тако да у узорку свакако важи $n < 30$) дата је формулом:

$$\left[\overline{x_n} - t\frac{s_n}{\sqrt{n}}\sqrt{1-\frac{n}{N}}, \overline{x_n} + t\frac{s_n}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\right]$$

где је t квартил студентове расподеле са $n-1$ степен слободе.

Интервал за ово обележје у случају $n = 3$ је $[192.354599850183, 193.645400149817]$ Фрагмент кода који описује овај процес дат је у наставку:

```
n = 3
N = 30
alpha = .95
xi = team.num.of.players %>%
sample_frac(0.1, replace = FALSE) %>%
pull(nrow)

xn.est = mean(xi)
t = qt(1-alpha/2, n-1)
c(xn.est - t * sd(xi) / sqrt(n) * sqrt(1 - n/N), xn.est + t * sd(
    xi) / sqrt(n) * sqrt(1 - n/N))
```

4 Оцењивање напредних статистика

У овом делу покушаћемо да оценимо неке напредније статистике које се воде у току ЕнБиЕј сезоне. Конкретно, плус-минус статиистику

и проценат акција у којима је одређени играч завршио напад свог тима било то покушајем шута, шутирањем слободног бацања или његовом грешком током утакмице.

4.1 Систематски узорак за оцењивање плус-минус статистике

Појам плус-минус статистика је појам који није глобално прихваћен, већ представља одомаћен назив за појам нет-рејтинг (енг. *net-rating*), због чињенице да је једина кошаркашка статистика која узима вредности из негативног дела скупа $\mathbb{Z}$. А представља кош разлику коју током утакмице има тим играча док је он у игри. Очекивано је да ова вредност на нивоу целе популације буде у близини нуле.

То ћемо и проверити систематским узорком, чиме обезбеђујемо да број играча у узорку из сваке сезоне буде сличан. Узорак се при оваквом начину узорковања добија тако што након што одредимо обим узорка рачунамо корак $k = \frac{N}{n}$, и након тога из списка јединца у оквиру за одабир бирамо сваку к-ту.

Оцена средње вредности обележја приликом систематског узорковања дата је формулом:

$$\widehat{x_{sis}} = \frac{1}{n} \sum_{i \in S} x_i$$

Док је оцена дисперзије те оцене дата са:

$$\widehat{D\left(\widehat{x_{sis}}\right)} = \frac{\overline{S}^2}{n}\left(1 - \frac{n}{N}\right)$$

где $\overline{S}^2$ представља узорачку дисперзију. Фрагмент Ар кода који оцењује тражену статиску систематским узорком дат је у наставку:

```
sys.sample.net.rating = function(n) {
    k = round(nrow(nbaPlayers)/n)
    net.rating = nbaPlayers$net_rating[seq(sample(k, 1), nrow(
     nbaPlayers), k)]
    cat('Ocena srednje vrednosti plus-minus statistike ', mean(net
     .rating), '\n')
    return(net.rating)
}
n = 30
net.rating = sys.sample.net.rating(n)
cat('Ocena disperzije ocene je: ', est.d.srswor(sampled.data = net
     .rating, n, nrow(nbaPlayers)), '\n')
```

Једном симулацијом фрагмента 4.1 за $n = 100$ добијамо вредност: -2.085149 и оцену дисперзије оцене: 1.003026. Што можемо сматрати релативно очекиваним резулататом.

4.2 Регресионо оцењивање статистике *usg_pct* на основу простог случајног узорка без понављања

У овом делу истраживања, желели смо да тестирамо конкретну статиску јер можемо искуствено да очекујемо приближан резултат. Ако знамо да је у сваком моменту пет играча на терену, очекивано је да средња вредност овог обележја буде ~ 0.2.

Оцена средње вредности обележја приликом регресионог оцењивања на основу простог случајног узорка дата је формулом:

$$\widehat{x_{LR}} = \widehat{x_n} + \hat{b}(\hat{y} - \hat{y_n})$$

. Као што се види из формуле изнад за ову анализу нам је потребна вредност популацијске статистике обележја који је у линеарној вези са обележјем од интереса. У овом оквиру за одабир обележје које највише има смисла да одговара овом опису је просечан број поена, јер што чешће играч завршава нападе свог тима то су већи изгледи да постигне поене. Популацијску статиску посечног броја поена преузет је са ове локације. Преузети су подаци о просечном броју поена на утакмици током тражених сезона и након тога је то скалирано на 12 играча. Парамтер $\hat{b}$ ће такође бити оцењен формулом:

$$\hat{b} = \hat{\rho}\frac{s_n(x)}{s_n(y)}$$

Где је ρ коефицијент корелације.

```
#source: https://www.basketball-reference.com/leagues/NBA_stats_
    per_game.html#stats::1
pts.ext.source = c(111.8, 111.2, 106.3,
                   105.6, 102.7, 100.0,
                   101.0, 98.1, 96.3,
                   99.6, 100.4, 100,
                   99.9, 98.7, 97,
                   97.2, 93.4, 95.1,
                   95.5, 94.8, 97.5,
                   91.6, 95.6, 96.9)
pts.pop = mean(pts.ext.source/12)
srswor.sample = (nbaPlayers %>% select(pts, usg_pct))[sample(nrow(
    nbaPlayers), 100, replace = FALSE),]
srswor.sample
rho.hat = cov(srswor.sample$pts, srswor.sample$usg_pct)
b.hat = rho.hat * sd(srswor.sample$usg_pct) / sd(srswor.sample$pts
    )
x.lr.hat = mean(srswor.sample$usg_pct) + b.hat * (pts.pop - mean(
    srswor.sample$usg_pct))
cat('Ocena srednje vrednosti obelezja usg_pct je ', x.lr.hat)
```

Вредност оцене коју добијемо за $n = 100$ је 0.2017308, што је управо оно што је било очекивано.

4.3 Количничко оцењивање напредних статистика везаних за ухваћене лопте

Напредне статистике које се односе на ухваћене лопте представљају обележја *oreb_pct* и *dreb_pct* у нашем скупу података. Њихова вредност означава проценат ухваћених лопти након промашаја од стране играча током времена проведеног у игри. *oreb_pct* представља ову вредност за напад тј. након промашаја тима у ком играч наступа и аналогно друго обележје за одбрану.

Ова статистика релевантније показује ангажовање играча у овом пољу кошаркашке игре у односу на обележје које говори просечан број скокова по мечу, јер минутажа играча није униформно распоређена. Желимо да оценимо средњу вредност ових обележја на популацији, са претпоставком да је очекивано да су обе статистике у близини вредности 0.1. Ако знамо да се у једном тренутку налази по пет играча

Оцењена средња вредност *oreb_pct*	0.06307307
Оцењена средња вредност *dreb_pct*	0.1493784

Табела 2: Оцене карактеристичне статистике за скок

овог тима евидентно је зашто се ово очекује, са мало већом вредношћу за обележје *dreb_pct*, због чињенице да играчи који бране кош се у већини случајева налазе ближе кошу и самим тим постоји већа вероватноћа да играчи тима који се брани ухвате лопту.

Као што се види у наслову ове подсекције, користићемо количничко оцењивање на основу узорковања јединки са неједнаким вероватноћама избора.

Оцена средње вредности у случају коришћења овог метода дата је формулом:

$$\overline{X_R} = \frac{\sum_{i=1}^{V} \frac{x_i}{\pi_i}}{\sum_{i=1}^{V} \frac{y_i}{\pi_i}} \overline{y}$$

где π_i представља вероватноћу укључења првог реда и-те јединице у узорку, а y_i и $\overline{y}$ редом вредност помоћног обележја на и-те јединице у узорку и популацијску средњу вредност тог обележја.

У нашем случају, помоћно обележје је висина играча, јер постоји јасна линеарна веза између висине кошаркаша и његове потентности при хватању лопте. Популацијску средњу вредност за обележје висина преузето је са спољног извора. [4]

Оцена дисперзије оцене средње вредности дата је формулом:

$$\widehat{D(\overline{X_R})} = \frac{\overline{y}^2}{t_y^2} \left[\sum_{i=1}^{V} \frac{1-\pi_i}{\pi_i^2} (x_i - \hat{R}y_i)^2 + \sum_{i=1}^{V} \sum_{\substack{j=1 \\ j \neq i}}^{V} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{(x_i - \hat{R}y_i)(x_y - \hat{R}y_j)}{\pi_{ij}} \right]$$

За формирање вероватноћа одабира у узорак коришћена је лажна претпоставка познавања тежине свих јединица у оквиру за одабир. Што је једино одступање од симулације реалне ситуације током овог мини истраживања.

Претпоставка се испоставља као тачна. Обе вредности су у близини 0.1, са већом вредности за одбрану из горе поменутих разлога. Закључак је да вероватноћа да играч који брани кош ухвати лопту након промашаја од играча који покушава да постигне кош два пут већа.

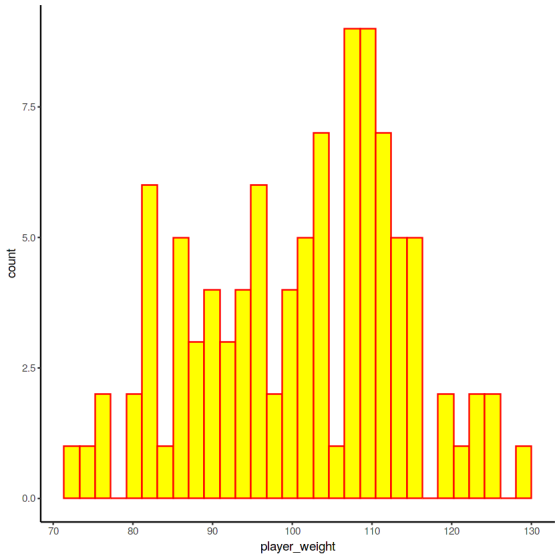
Фрагмент Ар кода који симулира цео описан процес дат је у наставку:

```
#source: https://www.thehoopsgeek.com/average-nba-height/
height = c(5+5/12, 5+9/12, 5+10/12, 5+11/12, 6 + (0:11)/12, 7 +
    (0:6) / 12) * 30.48
num.of.players = c(13, 21, 58, 90, 282, 423, 386, 688, 591,
663, 899, 1017, 968, 1344, 1013, 906, 595, 158, 80, 41, 6, 4, 17)
avg.height = sum(height * num.of.players) / sum(num.of.players)

sampled.weight.players = nbaPlayers %>%
sample_n(size = 100, replace = TRUE, weight = player_weight)

```

[4] Подаци са ове локације немају ажуриране податке за последње две године, па је претпостављено да ће остала 21 сезона одлично оценити популацијску вредност за наш скуп од 23 сезоне.



Слика 1: Хистограм обележја тежина играча над узарком за $n = 100$

```
ggplot(sampled.weight.players, aes(player_weight)) +
geom_histogram(color = "red", fill = "yellow", show.legend = TRUE)
      + theme_classic()


total.weight = sum(nbaPlayers$player_weight)
p = sampled.weight.players$player_weight / total.weight
n = 100
pi = 1- (1 - p)^n
R.hat.oreb = sum(sampled.weight.players$oreb_pct) / sum(sampled.
     weight.players$player_height)
R.hat.dreb = sum(sampled.weight.players$dreb_pct) / sum(sampled.
     weight.players$player_height)
oreb.mean.est = avg.height * R.hat.oreb
dreb.mean.est = avg.height * R.hat.dreb
cat(c(oreb.mean.est, dreb.mean.est))


pi.i.j = matrix(data = 0, nrow = n, ncol = n)

for (i in 1:n) {
    for (j in 1:n) {
        pi.i.j[i,j] = pi[i] + pi[j] - 1 + (1 - p[i] - p[j])^n

    }
}

est.D.oreb.mean.est = avg.height^2 /
            (sum(height * num.of.players)^2) +
            sum((1-pi) / pi^2 *
            (sampled.weight.players$oreb_pct - R.hat.oreb *
            sampled.weight.players$player_height)^2)

for (i in 1:n) {
    for (j in 1:n) {
        if (i != j) {
            est.D.oreb.mean.est =
            est.D.oreb.mean.est +
            (pi.i.j[i, j] - pi[i]*pi[j]) / (pi[i]*pi[j]) *
            (sampled.weight.players$oreb_pct[i] - R.hat.oreb *
            sampled.weight.players$player_height[i]) *
            (sampled.weight.players$oreb_pct[j] - R.hat.oreb *
```

```
            sampled.weight.players$player_height[j]) /
            pi.i.j[i, j]
        }
    }
}


est.D.dreb.mean.est = avg.height^2 /
        (sum(height * num.of.players)^2) +
        sum((1-pi) / pi^2 *
        (sampled.weight.players$dreb_pct - R.hat.dreb *
        sampled.weight.players$player_height)^2)

for (i in 1:n) {
    for (j in 1:n) {
        if (i != j) {
            est.D.dreb.mean.est = est.D.dreb.mean.est +
            (pi.i.j[i, j] - pi[i]*pi[j]) / (pi[i]*pi[j]) *
            (sampled.weight.players$dreb_pct[i] -
            R.hat.dreb * sampled.weight.players$player_height[i])
    *
            (sampled.weight.players$dreb_pct[j] - R.hat.dreb *
            sampled.weight.players$player_height[j]) /
            pi.i.j[i, j]
        }
    }
}


cat(est.D.oreb.mean.est, ' ', est.D.dreb.mean.est)
```

5 Оцене пропорција категоричких обележја

У овом одељку посветићемо се оцењивању пропорције до сад непоменутих обележја, као што су националност играча, колеџ који су завршили и то да ли су у лигу дошли на уобичајан начин путем избора (енг. *draft*) или су се прикључили лиги накнадно јер у почетку нису сматрани као довољно квалитетни.

5.1 Да ли је могућа врхунска кошарка и квалитетно образовање?

Неписано правило је да талентовани млади кошаркаши бирају слабије рангиране универзитете како би им остало више времена током колеџа које би користили за унапређивање својих кошаркашких вештина. Начин на који ћемо проверити пропорцију играча који долазе са престижних универзитета у лигу је тако што ћемо за престижан универзитет сматрати универзитет из листе Ајви лиге.

Метод којим ћемо оцењивати ову пропорцију је стратификован случајан узорак без понављања, где јединице стратификујемо на основу обележја *draft_number*. Пре тога ћемо елиминисати дуплиране јединице истог играча, како би елиминисали утицај броја година које играчи проводе у лиги. Што је мања вредност овог обележја значи да играч сматра талентованијим и вештијим (барем при доласку у лигу). Поделили смо оквир за одабир у стартуме, тако да су у истом стратуму јединице које имају исту вредност цифре десетица, док је постоји и додатни стратум за све јединице које нису учествовале на избору.

Оцена пропорције при овом методу узорковања је дата формулом:

$$\widehat{p_{str}} = \frac{1}{N}\sum_{h=1}^{L} N_h p_{n_h}$$

где је L број стратума, N обим оквира за узорковање, N_h величина сваког стратума, а p_{n_h} оцена пропорције у x-том стратуму.

Док је оцена дисперзије ове оцене дата са:

$$\widehat{D\left(\widehat{p_{str}}\right)} = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h - 1} p_{n_h}(1 - p_{n_h})$$

Фрагмент кода који симулира описан процес дат је у наставку:

```
ivy.league = c('Harvard', 'Cornell', 'Brown', 'Yale',
               'Dartmouth', 'Columbia', 'Princeton', 'Pennsylvania
    ')
all.players = nbaPlayers %>%
            group_by(player_name, college, draft_number) %>%
summarise(pts = mean(pts))
levels(all.players$draft_number) = c(levels(all.players$draft_
    number), 0)
all.players$draft_number[all.players$draft_number == 'Undrafted']
    = 0

all.players  = all.players %>%
mutate(draft.strat = ceiling(as.numeric(as.character(draft_number)
    ) / 10))

stratified.player.by.draft.num = all.players %>%
group_by(draft.strat) %>%
sample_frac(.1, replace = FALSE)


N = nrow(nbaPlayers)
N.h = (all.players %>%
       filter(draft.strat < 7) %>%
       group_by(draft.strat) %>%
       summarise(nrow = n()))$nrow


n.h = c()
A.n.h = c()
for (i in 0:6) {
    df = stratified.player.by.draft.num %>% filter(draft.strat ==
    i)
    n.h = c(n.h, nrow(df))
    A.n.h = c(A.n.h,nrow(df %>% filter(college %in% ivy.league)))
}
p.n.h = A.n.h / n.h


p.str = 1/N * sum(p.n.h * N.h)
D.p.str.hat = 1/N^2 * sum(N.h * (N.h - n.h) / (n.h - 1) * p.n.h *
    (1-p.n.h))

cat('Ocenjena proporcija strat uzorkom: ', p.str, '\n')
cat('Ocena disperzije ocene: ', D.p.str.hat)
```

Оцењена пропорција страт. узорковањем	0.0008972633
Оцена дисперзије оцене	0.0000007245734

Табела 3: Резултати добијени оцењивањем пропорције играча из Ајви лиге

Као што је и очекивано, а на основу резултата 5 бивши алумни колеџа из Ајви лиге нису претерано чести учесници националне кошаркашке лиге.

Знајући ову инфорамцију, тј. да је број ових играча мали. Можемо додатно погледати просечан број поена тих играча, да видимо како се рангирају у односу на остале не играче.

	player_name	college	draft_number	avg_pts_all_time
1	Chris Dudley	Yale	75	1.89
2	Ira Bowman	Pennsylvania	0	1.27
3	Jeff Foote	Cornell	0	1
4	Jeremy Lin	Harvard	0	12.0
5	Jerome Allen	Pennsylvania	49	3
6	Matt Maloney	Pennsylvania	0	5.7
7	Miye Oni	Yale	58	0
8	Steve Goodrich	Princeton	0	1.1

Табела 4: Алумни колеџа Ајви лиге у ЕнБиЕј лиги у периоду 1996-2020

Као што видимо у 4 већина играча није ни било изабрано на почетку, већ је накнадно примљено у лигу. Док вредност просечног броја поена од 3.24543650793651 говори да су значајно испод просека свих играча у лиги.

Фрагмент кода који нам пружа претходне информације:

```
mean((all.players %>% filter(college %in% ivy.league))$pts)
print(all.players %>% filter(college %in% ivy.league))
```

Све наведено горе недвосмислено говори да су играчи који одустају од озбиљне академске каријe зарад професионалног спорта јесу у праву.

5.2 Да ли је ЕнБиЕј лига затворена за неамеричке играче?

За ЕнБиЕј лигу важи да је неевропским играчима теже да се домогну места у њој него домаћим играчима. Зато ћемо простим случајним узорком са понављањем оценити пропорцију играча који немају америчко држављанство. Оцена пропорције овом методом дата је са:

$$p_n = \frac{a_n}{n}$$

где a_n представља број јединица у узорку које задовољавају тражено својство. Док оцена диспрезије ове оцене се може израчунати формулом:

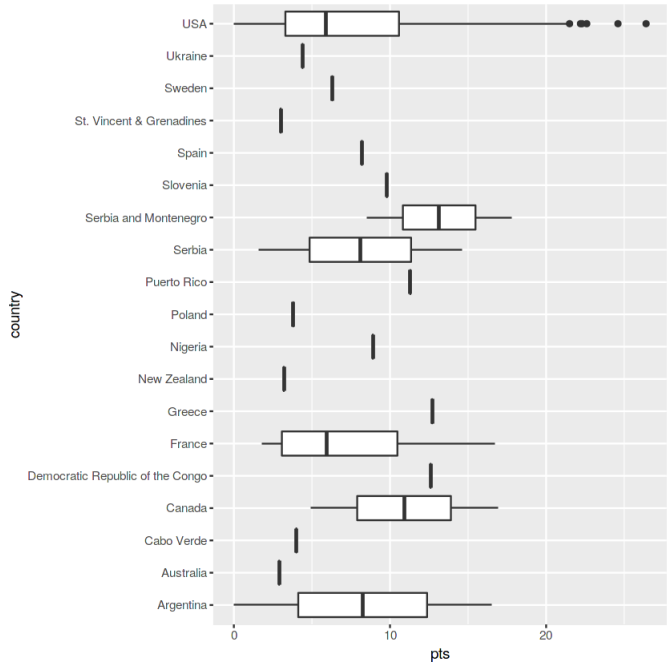
$$\widehat{D\left(\widehat{p_n}\right)} = \frac{p_n(1-p_n)}{n-1}$$

Фрагмент кода који симулира описану процедуру дата је у наставку:

```
n = 200
srswr.sample = nbaPlayers %>%
sample_n(size = n, replace = T) %>%
select(player_name, pts, country)
```

Оцењена пропорција псусп узорковањем	0.13
Оцена дисперзије оцене	0.0005683417
90-одсто процентна интервална оцена	(0.1245, 0.1354)

Табела 5: Оцена размере играча који нису америчке националности



Слика 2: Кутијасти дијаграм на узорку

```
a.n = srswr.sample %>%
group_by(country) %>%
summarise(nrow = n()) %>%
filter(country != 'USA') %>%
summarise(t = sum(nrow))

p.n = as.numeric(a.n) / n
cat('Ocena proporcije neamerkickih igraca: ', p.n , '\n')
cat('Ocena disperzije ocene: ', p.n * (1 - p.n) / (n - 1))

alpha = 0.90
z = qnorm(1-alpha/2)
c(p.n - z * sqrt((N-n) / N / (n-1) * p.n * (1-p.n)) - 1/(2*n),
  p.n + z * sqrt((N-n) / N / (n-1) * p.n * (1-p.n)) + 1/(2*n))
```

Имајући у виду и додатне податке из узорка (у нашем случају $n = 200$) 2 видимо да просечан број поена играча из других држава у већем броју превазилази исту статистику за играче из Америке показује нам да чињеница да сваки десети играч није амерички држављанин јасно показује да је ЕнБиЕј лига и даље предност даје домаћим играчима.

6 Закључак

Анализа узорака која је спроведена у претходним секцијама дала нам је оцену нумеричке вредности колика је предност при скоку играча у одбрани у односу на играче у нападу. Затим, потврдила нам је неке тезе које се искуствено могу претпоставити пуким праћењем ЕнБиЕј лиге, као нпр. да страни играчи имају тежи пут до лиге од америчкх играча, да је тешко подједнако се добро посветити академској и кошаркашкој каријери, као и да се играчи све више посвећују што квалитетнијим личним статистикама. Јер експанзијом технологија све више се улаже у праћење шареноликих параметара кошаркашке игре и воде се евиденције о најситнијим детаљима. У томе предњачи управо ЕнБиЕј лига, а константним изношењем таквих података у јавност самим играчима се ствара притисак да све те параметре поправљају. Занемарујући главни циљ игре, а то је победа тима.

Литература

[1] Sharon L. Lohr. *Sampling: Design and Analysis*. Brooks/Cole, Cengage Learning, 2010.

[2] доц. др Ленка Главаш. Материјали са предавања Увод у теорију узорака, 2020. интернет локација: http://www.matf.bg.ac.rs/p/lenka-zivadinovic/kurs/687/%D0%A3%D0%B2%D0%BE%D0%B4-%D1%83-%D1%82%D0%B5%D0%BE%D1%80%D0%B8%D1%98%D1%83-%D1%83%D0%B7%D0%BE%D1%80%D0%B0%D0%BA%D0%B0-4%D0%98/.

A Додатак

- Интерактивно окружење са могућношћу извршавања свих симулација поменутих у раду могуће је видети овде.