UNIVERSITETI I PRISHTINËS
FAKULTETI I SHKENCAVE MATEMATIKORE DHE NATYRORE
DEPARTAMENTI I MATEMATIKËS

Punim Seminarik në Statistikë

Tema: F-testi me anë të gjuhës R

Mentori: **Studentët:**

Prof. Edmond Aliaga **Donat Sherifi, Getoar Krasniqi & Artin Hajdari**

Prishtinë-/2021/01/10/

Informata bazike rreth gjuhës R

R është një gjuhë dhe ambient për llogaritje statistikore dhe grafike. Është një projekt GNU (Operating System), i cili është i ngjashëm me gjuhën dhe ambientin S (i zhvilluar tek Bell Laboratories (ish AT&T, tani Lucent Technologies) nga John Chambers me kolegë). R mund të konsiderohet si një implementim më ndryshe i S. Janë disa ndryshime të rëndësishme, por shumica e kodit për S ekzekutohet pa ndonjë problem edhe në R.

R jap një shumëllojshmëri të gjerë të teknikave statistikore dhe grafike, dhe është shumë i zgjerueshëm. Gjuha S është zgjedhja më e shpeshtë për kërkime në metodologjinë statistikore, dhe R siguron një rrugë Open Source për të marrë pjesë në këto aktivitete.

Njëra ndër fuqitë e R është lehtësia me të cilën mund të prodhohen pjesë të dizajnuara mirë dhe kualitative, dukë përfshirë edhe simbole e formula matematikore aty ku paraqitet nevoja. Është marrë një kujdes i mrekullueshëm për zgjedhjet e vogla dizajnuese grafikore, por në fund të fundit shfrytëzuesi është nën kontrollë të plotë rreth dukjes së ambientin (environment).

Ambienti R përfshinë:

- një sistem efektiv për ruajtje dhe manipulim me të dhëna,
- një komplet të operatorëve për kalkulime në vargje, posaqërisht në matrica,
- një koleksion të madh, koherent dhe të integruar veglash për analiza të të dhënave,
- sisteme grafike për analizë të të dhënave, shfaqje në ekran ose në kopje fizike, dhe
- një gjuhë programuese të zhvilluar mirë, të thjeshtë, e cila përmbanë kushtëzore (conditionals), unaza (loops), funksione rekurzive të japura nga shfrytëzuesi dhe sisteme hyrëse dhe dalëse.

F-Testi

‘**Testi F’** është një test në fushën e statistikës me rëndësi mjaft të madhe i cili neve na tregon nëqoftëse modeli ynë i të dhënave ofron një përshtatje më të mirë të të dhënave sesa në krahasim me një model që nuk përmban të dhëna të varura nga populacioni. Ky model më së shpeshti përdoret kur krahasohen modelet e ndryshme statistikore brenda me një grup të dhënash, në mënyrë që të identifikojë modelin që i përshtatet më së miri popullatës nga e cila janë marrë të dhënat në formë të çiftit të mostrave.

Për ta kuptuar më mirë Testin ‘F’ në fushën e statistikës, ne do ta tregojmë matematikën prapa këtij testimi statistikor. Staticientët me kalimin e kohës zbuluan se kur çiftet e mostrave merren nga një popullatë normale, raportet e variansave të mostrave në secilën palë do të ndjekin gjithmonë të njëjtën shpërndarje, kjo shpërndarje quhet shpërndarja F. Nëse të dy variancat janë

shumë afër të qenit të barabartë pra vlerës ‘1’, atëhere dy mostrat vijnë lehtë nga popullata të barabarta. Matematikisht llogaritet si raporti i variancave të mostrave përkatëse me formulën në vijim:

$$F = s_1^2 / s_2^2$$

Tani vlerat ‘s1’ dhe ‘s2’ paraqesin variansat për dy mostrat përkatëse nga populacioni. Ato llogariten si shuma e ndryshimit të katrorëve me mesatare thyer për madhësinë e mostrës minus një, si me formulën në vijim:

$$s^2 = \sum (x - \bar{x})^2 / (n - 1)$$

Pasi që kemi llogaritur variansat ne jemi në gjendje të llogaritim vlerën ‘F’. Pastaj caktojmë hipotezat për testim, ato janë ‘Ho’ kur variansat i takojnë populacionit të njëjtë dhe ‘Ha’ kur nuk i takojnë, me formulat në vijim:

$$H_o : \sigma_W^2 = \sigma_S^2$$

$$H_a : \sigma_W^2 \neq \sigma_S^2$$

Në vijim pas këtyre hapave ne duhet ta përdorim “Tabelën F” për ta caktuar më saktë krahasimin midis vlerës së llogaritur ‘F’ dhe vlerës kritike‘F’ për testim të hipotezës. Fillimisht për ta përdorur këtë tabelë ne duhet ti dimë vlerat e shtyllës dhe rreshtit. Pasi që tek testimi F ne në raport të vlerës kritike e paraqesim variansën më të madhe në numërues dhe më të voglën në emërues atëherë, shtyllat do të paraqesin ‘Shkallët e lirisë për Numërues’ ndërsa rreshtat ‘Shkallët e lirisë për Emërues’. Shkalla e lirisë llogaritet lehtë duke e kalkuluar madhësin e mostrës përkatëse minus një. Pas kësaj neve na nevojitet edhe vlera ‘Alfa’ për testim të cilën ne e caktojmë që të jetë në vlerën: $\alpha = 0.05$. Kjo ka vlerë mjaft të rëndësishme pasi që na dikton neve tabelën ‘F’ që do të përdorim për testim. Ne përdorim tabelën ‘F’ për $\alpha = 0.05$, si në vijim:

F Table for alpha=.05

DF1 (see next table below for 12 to ∞)

DF2	1	2	3	4	5	6	7	8	9	10
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782

Nga kjo tabelë e nxjerrim vlerën kritike ‘F’ të cilën e krahasojmë me vlerën e llogaritur ‘F’. Nëse vlera e llogaritur është më e madhe sesa vlera kritike atëherë hipoteza ‘Ho’ bie.

F-Testi në gjuhën R

Gjuha R është gjuhë që përmbanë vegla të shumta të automatizuara për matje dhe testime të ndryshme statistikore. Kjo gjuhë përmbanë elemente mjaft të ngjajshme dhe funksionalitetet bazike të gjuhëve programuese të tjera siç janë variablat, vektorët, funksionet etj. Gjuha R përmbanë veglat e nevojshme edhe për krahasime të variancave të popullimeve të ndryshme ku shfrytëzohet F-Testi. Si funksion kryesor që shërben për krahasim mes variancave të dy mostrave që mund të jenë të madhësive të ndryshme është funksioni var.test() i cili merr disa parametra dhe shfrytëzon F-Testin për të arritur tek rezultati. Për këto testime ne kemi marrë datasetin e gatshëm nga gjuha R, ToothGrown. Gjuha R në vete përmbanë edhe datasete tjera testuese që mund të shfrytëzohen për analizime të ndryshme. Disa nga parametrat kryesorë të funksionit var.test () ne mund t’i shohim në vijim:

```
# S3 method for default
var.test(x, y, ratio = 1,
         alternative = c("two.sided", "less", "greater"),
         conf.level = 0.95, …)
```

Nga figura e mësipërme vërejmë se në shumicën e rasteve funksioni var.test() merr si parametër dy vektorë ku si të dhëna në këta vektorë duhet të jenë frekuencat e karakteristikës së matur në dy

popullimet. Variabla ratio tregon për proporcionin e variancave të cilin e supozon hipoteza Ho. Pastaj ekziston mundësia e ndërrimit të llojit të testit nëpërmjet variablës alternative ku janë disa mundësi. Opsioni "two.sided" tregon për testim të dyanshëm, pastaj kemi opsionin less për bishtin e djathtë dhe greater për bishtin e majtë të shpërndarjes. Gjithashtu ekziston edhe mundësia e ndërrimit të nivelit të besueshmërisë së testit nëpërmjet variablës conf.level e cila gjithashtu e cakton edhe nivelin e signifikancës alfa.

Në kodin e mëposhtëm është marrë dataseti nga paketat e gjuhës R i cili përmbanë 60 rekorde të gjatësive të dhëmbëve të disa kafshëve të një lloji pas trajtimit me dy lloje të ndryshme suplementesh. Më poshtë mund të shohim kodin që bën ngarkimin e datasetit si dhe paraqitjen e një mostre 10-elementëshe nga ky dataset:

```
> install.packages("dplyr")

 > testData = ToothGrowth
 > library("dplyr")

> sample_n(testData, 10)
    len supp dose
1  21.5   VC  2.0
2   4.2   VC  0.5
3  26.4   OJ  2.0
4  11.2   VC  0.5
5   6.4   VC  0.5
6   9.7   OJ  0.5
7  16.5   VC  1.0
8  30.9   OJ  2.0
9  15.2   OJ  0.5
10 16.5   OJ  0.5
>
```

Shohim se meqë duam të paraqesim një mostër 10-elementëshe na duhet instalimi dhe përdorimi i paketës "dplyr". Pastaj me anë të funksionit sample_n() që merr parametër datasetin tonë dhe madhësinë e mostrës që duam ta shfaqim, ne shohim disa të dhëna për datasetin. Në kolonën "len" janë gjatësitë përkatëse dhe në kolonën "supp" janë suplementet që janë përdorur. Kolona dose tregon për dozat, por këto nuk do të na interesojnë gjatë testimit tonë. Tani me anë të kodit të mëposhtëm ne bëjmë krahasimin e variancave të gjatësive nga njëri suplement me tjetrin:

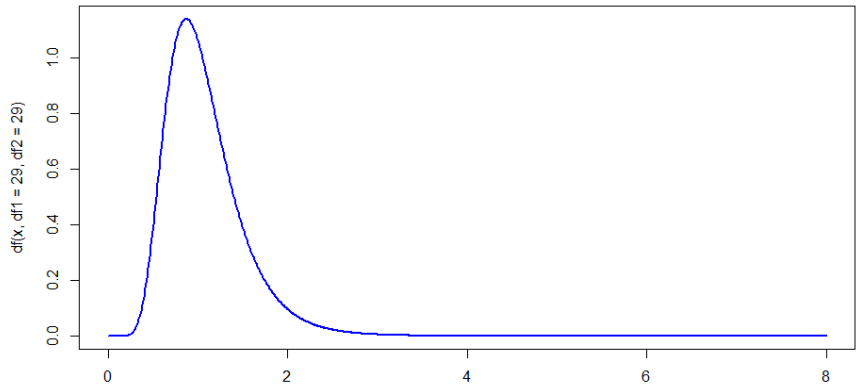
```
> var.test(len ~ supp, data = testData)

        F test to compare two variances

data:  len by supp
F = 0.6386, num df = 29, denom df = 29, p-value = 0.2331
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3039488 1.3416857
sample estimates:
ratio of variances
         0.6385951
```

Në kodin e mësipërm shihen disa parametra tjerë në funksionin var.test. Parametri i parë i tregon që frekuencat do të jenë nga kolona len ndërsa këto të dhëna grupohen sipas kolonës supp. Në parametrin e dytë vendoset tabela. Pas ekzekutimit na shfaqen rezultatet ku vërehen vlera F,

shkallët e lirisë për numërues, emërues si dhe vlera p që është më e madhe se alfa që është 0.025. Gjithashtu poshtë informacionit për intervalin e besueshmërisë kemi dy vlerat kritike pasi që ky test ka qenë i dyanshëm. Nga ky test ne vërejmë se nuk ka mjaft evidencë që rezultatet e njërit suplement janë më variabile se tjetrit.

Në figurën e mëposhtme shohim kodin dhe paraqitjen grafike të shpërndarjes F për shkallët e lirive të dhëna nga mostrat tona:

```
> curve(df(x, df1 = 29, df2 = 29), from=0, to=8, n=1000, col= 'blue', lwd=2)
```



Në teorinë e gjasës dhe statistikë, **shpërndarja F** (po ashtu e njohur si **F-shpërndarja e Snedecor**-it ose **shpërndarja e Fisher-Snedecor**) është një shpërndarje e vazhdueshme e gjasës e cila njihet shpesh si shpërndarja 'null' e një testi statistikor, më së shumti në analizën e variancës (ANOVA), p.sh. F-Testi.

Nëse një ndryshore e rastësishme X ka një shpërndarje F me parametra *d1* dhe *d2*, ne shkruajmë X~F(d1,d2). Atëherë funksioni i dendësisë së gjasës për X është i japur me:

$$f(x;d_1,d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1}\ d_2^{d_2}}{(d_1 x+d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2},\frac{d_2}{2}\right)}$$

$$= \frac{1}{\mathrm{B}\left(\frac{d_1}{2},\frac{d_2}{2}\right)}\left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1}\left(1+\frac{d_1}{d_2}\,x\right)^{-\frac{d_1+d_2}{2}}$$

për çdo x>0. Këtu B është funksioni beta. Në shumë aplikime, parametrat *d1* dhe *d2* janë numra të plotë pozitivë, por shpërndarja është mirë e definuar edhe për vlera reale pozitive të këtyre parametrave.