جزوه جلسه هشتم دوره آمار با R دکتر مجید صابری

گردآورنده: مرضیه رحیمی، بهار 2024

مروری بر کلیات جلسه قبل:

دو جلسه قبل در مورد مقایسه­های آماری و دو گروهی و جلسه قبل در مورد مقایسه چند گروهی(بیشتر سه گروهی) صحبت کردیم. کلیت نگاه ما این بود که اول ببینیم می­خواهیم از آمار پارامتریک و یا ناپارامتریک استفاده کنیم. بر اساس اینکه متغیر وابسته ما چه تابع توزیعی دارد؛ عمل می­کردیم که اگر تابع توزیع یکی از این­ها نرمال نبود، ما باید به اجبار به سراغ آمار ناپارامتریک برویم. در قسمت پایانی کارمان در جلسه قبل بررسی کردیم که بر اساس اینکه چند متغیر مستقل داریم میتواند تست ما یک راهه[[1]](#footnote-1) دو راهه یا سه راهه باشد، اما متغیر وابسته­مان همچنان یکتا هست.

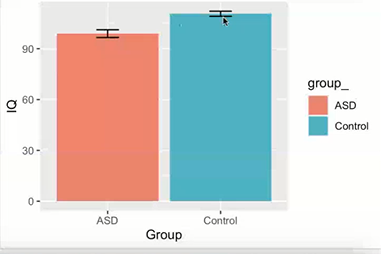
مثلا در مثال جلسه قبل؛ دو متغیر مستقل (گروه(سالم یا کنترل) و سایت(مکان جمع آوری دیتا)) و یک متغیر وابسته (FIQ) داریم. چون دو متغیر مستقل داریم می­شود two way anova. درجه این anova هم بر اساس تعداد levelهای ما مشخص می­شود. گروه ما دوتا بود؛ اتیسم و سالم و سایت ما سه تا بود؛ پس یک two way anova داریم که 2\*3 هست(دو سطح برای گروه داریم و سه سطح برای سایت). بعد از زدن two way anova در سطح پایین­تر هر کدام را یک one way anova می­زنیم، در واقع یعنی به دو انووای یک راهه شکسته می­شود و تست­های post hoc را برای آن ها انجام می­دهیم(سایت جداگانه و گروه جداگانه) در واقع انگار اثر هر کدام را به صورت جداگانه بررسی می­کنیم (فارغ از گروه، ببینیم آیا تفاوتی بین جاهای مختلفی که دیتا گرفته شده وجود دارد و برعکس؛ یعنی آیا بدون در نظر گرفتن تفاوت­های سایت­ها بین دو گروه ما تفاوتی وجود دارد یا نه). گام بعدی آزمون­های post hoc هستند که دو به دو گروه­ها با هم مقایسه شوند، و ما از توکی استفاده می­کنیم. در واقع وقتی در main effect ، p.value ما معنادار باشد به عنوان گام بعدی به سراغ این تست­ها می­رویم که ببینیم دقیقا آن تفاوت در کجا و بین کدام گروه­ها هست.

همین روند را می­توانیم به صورت ناپارامتریک داشته باشیم. متغیری که در ناپارامتریک استفاده می­کنیم در واقع ترتیبی[[2]](#footnote-2) هست و آن ها را رتبه بندی می­کنیم و بر اساس همدیگر رتبه بندی می­کنیم و تست ناپارامتریک می­زنیم.

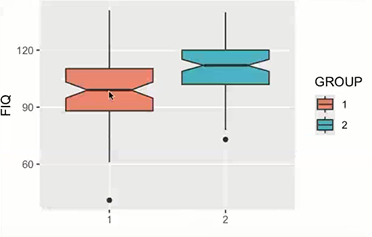
مورد دیگری که صحبت کردیم (در جدول خلاصه شده) measureهایی است که در هر مورد استفاده می­کنیم.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Index of central tendency** | **Index of dispersion** | **plot** |
| پارامتریک | میانگین | **Standard** **deviation** | **Bar plot** |
| ناپارامتریک | میانه | **Interquartile range** | **Box plot** |

به ازای هر main effect می­توانیم یک پلات داشته باشیم که بسته به موقعیت Bar plot یا Box plot هست. در Bar plot ارتفاع bar میانگین گروه هست و آن چیزی که بالای بار[[3]](#footnote-3) داریم در واقع Standard deviation هست که اگر با Standard deviation بار دیگر هم پوشانی نداشته باشد(که در این مثال ندارد) یعنی تفاوت معنادار در دو گروه می­بینیم.



در Box plot خط وسط میانه هست، خط پایین Q1 و خط بالا Q3 هست. ناحیه ای که با آکولاد مشخص شده Interquartile range هست. و آن ناحیه notch ، confidence intervalای هست که ما برای میانه تخمین زدیم(با دایره قرمز مشخص شده). اگر این نواحی برای دو گروه هم پوشانی نداشته باشد p.value معنادار هست.



\*اگر ما تابع توزیع­مان نرمال بود، می­توانیم هم از پارامتریک و هم از ناپارامتریک استفاده کنیم اما چون پارامتریک دقیق­تر هست و استفاده از ناپارامتریک در بسیاری از مواقع صحیح نیست(چون در واقع از مقادیر ترتیبی استفاده می­کند) پس از پارامتریک استفاده می­کنیم اما اگر توزیع­مان نرمال نباشد حتما باید از ناپارامتریک استفاده کنیم.

جلسه هشتم(پایان مرور و شروع مطالب جدید):

با هم پیش بریم و کدها رو ران کنیم. اول فضای کار را پاک می­کنیم و library مد نظر رو فراخوانی می­کنیم.

##Clearing workspace

rm(list = ls())

##Loading required packages

library(ggplot2)

##Setting the path

setwd("") #set to your path

##IQ Sample

pheno <- read.csv("Phenotypic.csv")

pheno[pheno == "-9999"] <- NA

####NYU site

pheno\_NYU <- pheno[pheno$SITE\_ID == "NYU" ,]

tabulate(as.factor(pheno\_NYU$DX\_GROUP))

امروز هم فقط با دیتاست NYU کار داریم و پیش می­رویم.

در واقع می­خواهیم دو measure ، PIQ و VIQ را برداریم و یک جدول از رویشان بسازیم.

###FQI for NYU site

VIQ\_Control <- pheno\_NYU$VIQ[pheno\_NYU$DX\_GROUP == 2]

VIQ افراد سالم

VIQ\_ASD <- pheno\_NYU$VIQ[pheno\_NYU$DX\_GROUP == 1]

VIQ افراد اتیسم

PIQ\_Control <- pheno\_NYU$PIQ[pheno\_NYU$DX\_GROUP == 2]

PIQ افراد سالم

PIQ\_ASD <- pheno\_NYU$PIQ[pheno\_NYU$DX\_GROUP == 1]

PIQ افراد اتیسم

اول بردارهایمان را از این مقادیری که می­خواهیم را می­سازیم؛

data <- data.frame(c(pheno\_NYU$SUB\_ID,pheno\_NYU$SUB\_ID),ستون اول

c(rep("VIQ",184),rep("PIQ",184)),ستون دوم

c(rep("Cont",105),rep("ASD",79),rep("Cont",105),rep("ASD",79)),ستون سوم

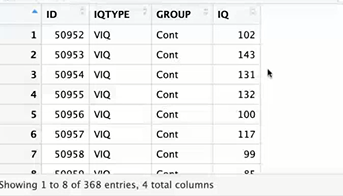
c(VIQ\_Control,VIQ\_ASD,PIQ\_Control,PIQ\_ASD))ستون چهارم

در این خط می­خواهم که دوبار این 184تا برای من تکرار شوند که یک بار VIQ و بار دیگر PIQ باشد.

colnames(data) <- c("ID","IQTYPE","GROUP","IQ")

این خط هم برای ست کردن اسم ستون های ما هست.

خروجی این کد جدول ما هست. بخشی از جدول:



در واقع چهار ستون هست که id فرد، iq type، گروه و iq هست.

در جلسه قبل در مورد انووا صحبت کردیم و گفتیم که در واقع اتفاقی که می افتد، دو یا سه گروه داشتیم برای مقایسه اما؛ فکر کنید یک گروه داریم که repeated measure هست. در این مورد من دوتا measurement دارم از یک سری افراد ثابت پس این repeated measure هست. اینجا متغیر وابسته من iq و متغیر مستقل iq type هست.

##One way repeated measure test

#Parametric

model <- aov(IQ ~ IQTYPE , data = data)

summary(model)

model <- aov(IQ ~ IQTYPE + Error(ID/IQTYPE), data = data)

summary(model)

قسمت هایلایت شده به این منظور هست که ما بگوییم id را در نظر بگیر و ببین چطور هست تفاوت iqها. اگر نه پیش فرض این هست که اینها ادم های متفاوتی هستند در حالی که اینجا ما یک سری ادم ثابت با دو نوع نمره iq داریم و اینطور هر subject را با خودش مقایسه می­کند.

معادل ناپارامتریک، فریدمن تست هست. آن مشخصه­ای که به ما می­گوید که سابجکت ها چطور هست در این ناپارامتریک این طور هست.

#Non-parametric

model <- friedman.test(IQ ~ IQTYPE | ID, data = data)

model

الان چون ما همین یک مورد را هم پارامتریک زدیم و هم ناپارامتریک. در حالی که پارامتریک معنادار شد و ناپارامتریک معنادار نشد. این به ما اهمیت انتخاب درست امار پارامتریک و ناپارامتریک را یادآوری می­کند.

وقتی که repeated measure داریم نمودارهایی به شکل زیر خواهیم داشت.

#Plotting

ggplot(data, aes(x=IQTYPE, y=IQ, group=ID)) +

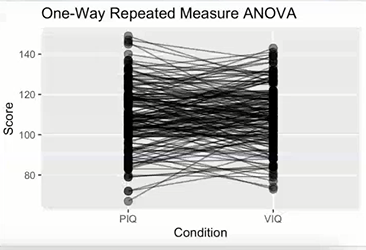
geom\_line(alpha=.5) +

geom\_point(size=3,alpha=.5 ) +

labs(title="One-Way Repeated Measure ANOVA",

x="Condition",

y="Score")



در واقع هر نقطه در هر سمت جایگاه فرد را در آن محور مشخص می­کند و این خطی که به هم وصل می­شود یعنی یک سابجکت را نشان می­دهد. ما از روی نمودار می­توان در مورد نتیجه تست اظهار نظر کرد. چطور؟ در واقع اگر معنادار باشد و p.value خوبی باشد در واقع اکثریت خط­های مورب ما در یک جهت خواهند بود، مثلا اگر PIQ از VIQ بزرگتر باشد عموم خط­ها به سمت محور VIQ حالت نزولی خواهند داشت و یک روند را به خوبی میتوان مشاهده کرد. این شکل خیلی p.value معناداری نداریم برای همین نمیتوان یک ترند مشخص را مشاهده کرد.

با کد زیر رنگ را می­توان به نمودارمان اضافه کرد که در واقع برای دو گروه اتیسم و عادی دو رنگ انتخاب می­کند و شاید اینطور برای گروه ها میتوان یک ترند را دید.

ggplot(data, aes(x=IQTYPE, y=IQ, group=ID, color=GROUP)) +

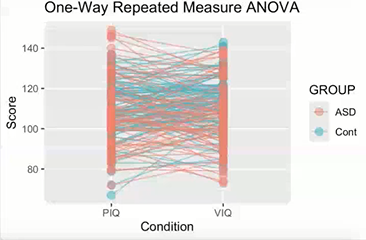
geom\_line(alpha=.5) +

geom\_point(size=3,alpha=.5 ) +

labs(title="One-Way Repeated Measure ANOVA",

x="Condition",

y="Score")



\*تکلیف:

یک دفعه افراد سالم را در نظر بگیرید و یک دفعه افراد بیمار را به صورت جداگانه همین کار را انجام بدهید (همین نمودار و تست های آماری؛ دو تا one way repeated measure) و از نظر آماری گزارش کنید که وقتی پارامتریک نگاه می­کنیم و وقتی ناپارامتریک باشد، چه تفاوت­هایی خواهد داشت و این p.valueای که انووا به ما می­دهد آیا جهت را هم به ما می­دهد یا باید توکی بزنیم تا بفهمیم جهت چطور هست؟

ادامه جلسه:

برویم سراغ Interaction effects؛ وقتی که یک راهه هست و یک متغیر مستقل داریم که دیدیم. اگر دو تا باشد بین متغیرهای مستقل + می­گذاریم.

#Main effects

model <- aov(IQ ~ IQTYPE + GROUP , data = data)

summary(model)

این بررسی این را به ما می­گوید که آیا IQ روی IQTYPE و GROUP تاثیر دارد یا نه.

بررسی این دو مورد به اضافه اینکه آیا interaction گروه و IQTYPE با هم روی IQتاثیر دارد یا نه را با این کد زیر میسنجیم.

#Main effects + interaction effect

model <- aov(IQ ~ IQTYPE \* GROUP , data = data)

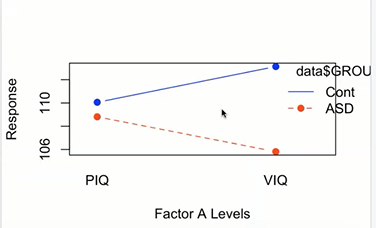
summary(model)

یا اینکه به این مدل بنویسیم:

model <- aov(IQ ~ IQTYPE \* GROUP + IQTYPE:GROUP, data = data)

summary(model)

نمودار interaction plot را اگر بکشیم؛



اگر این نمودار دو خطی، متقاطع باشد یعنی interaction ما مهم است یعنی PIQ و VIQ ما با هم رابطه مستقیم و یا معکوس را دارند، اما اگر متقاطع نباشد یعنی که رابطه معناداری بین آنها وجود ندارد و اگر موازی باشند هم معنادار نیست. در این مورد این دو خط هم را قطع نمی­کنند پس رابطه ما معنادار نیست.

ما وقتی یک آنووای دو راهه میزنیم محتوایی که از ان می­توانیم بیرون بکشیم Main effects هست و interaction ­ها. می­توانیم بگوییم که Main effects را کنار بگذار و فقط در مورد interaction بگو و جواب جدیدی به ما می­دهد که این یکی معنادار هست.

#Interaction effect

model <- aov(IQ ~ IQTYPE:GROUP, data = data)

summary(model)

چرا متفاوت هست؟ چون این دفعه روی interaction متمرکز می­شود و با پیش فرض اینکه هیچ Main effects ای وجود ندارد و فقط در مورد interaction محاسبات انجام می­دهد و در این مورد جواب دقیق­تری به ما می­دهد.

#Two-way repeated measure

model <- aov(IQ ~ IQTYPE \* GROUP + Error(ID/IQTYPE \* GROUP), data = data)

summary(model)

چون که ما دو measure از یک سری فرد داریم پس باید repeated measure بزنیم. و اینجا interaction را که میبینیم معنادار نیست و در این کد بالا این مورد را اصلاح کردیم.

کد نمودار(نمودار قبلی) نیز به این صورت خواهد بود:

#Ploting

interaction.plot(x.factor=data$IQTYPE, trace.factor=data$GROUP, response=data$IQ,

type="b",

pch=19,

col=c("red", "blue"),

xlab="Factor A Levels", ylab="Response", legend=TRUE)

1. One way [↑](#footnote-ref-1)
2. ordinal [↑](#footnote-ref-2)
3. Bar ستون [↑](#footnote-ref-3)