png(file="C NAFL NASH方差分析有差异.png",width=2500,height=2500)

par(mfrow=c(9,9))

for (aim in rownum2){

rect\_frame<-data.frame(intensity=c(as.numeric(A[aim,]),as.numeric(B[aim,]),as.numeric(C[aim,]),as.numeric(D[aim,]),as.numeric(E[aim,]),as.numeric(FF[aim,])),level=rep(c("A","B","C","D","E","F"),c(7,7,7,7,7,7)))

boxplot(tapply(rect\_frame$intensity, rect\_frame$level, list), main = na\_data[aim,2],varwidth = T,ylab = 'intensity',cex.main=3,cex.axis=2,cex.lab=2)

}

dev.off() #关闭右下角预览 相反是dev.new()

ggsave("a.png") 直接保存进工作目录

boxplot(tapply(df\_lm$肝脏脂肪含量, df\_lm$地理分区1南方2北方, list), main = "南北方人群肝脏脂肪含量箱线图",varwidth = T,border = c("black","red"),names = c("南方", "北方"), ylab = '肝脏脂肪含量(log10)')

带全套标注的boxplot

boxplot(tapply(dataframe[,t],dataframe$fibrosis,list),varwidth = F,

names=paste(c("F0,n=","F1,n=","F2,n=","F3,n=","F4,n="),table(dataframe[is.na(dataframe[,t])==FALSE,]$fibrosis)[1:5]),

main=paste(colnames(dataframe)[t]," P=",p),cex.axis=0.6) #axis坐标轴字大小

pheatmap(result,scale="row",cluster\_rows = T,cluster\_cols = F

,color = colorRampPalette(colors = c("green","black","red"))(100),na\_col = "black",border\_color=NA)

barplot(apply(数据集,1,mean)) 以行或列作柱状图（均值或中位数）

pie3D(入组受试者类型1正常2IGR3DM\_B超脂肪肝有无[1, ],

labels= paste(c("无脂肪肝", "脂肪肝"), " ",

round(入组受试者类型1正常2IGR3DM\_B超脂肪肝有无[1, ]/sum(入组受试者类型1正常2IGR3DM\_B超脂肪肝有无[1, ]), 4)\*100, "%", sep = "" ),

ggplot2：

qplot(log10(fpg),bmi,data=liverdata, color=sex) + geom\_smooth(method = lm) 以sex为分组散点图 先写X再写Y

color可以等于连续变量，出来渐变色

color改变线 fill改变填充

qplot(sex,log10(fpg),data=liverdata, geom="boxplot")

geom=bar density line 可作相应图

,alpha=0.01) 参数 点为半透明

上文qplot语句=

ggplot(liverdata,aes(bmi,log10(fpg),color=sex))+geom\_point()+ geom\_smooth(method = lm)

aes里先写X再写Y

color=sex：不同性别是不同色

shape=sex

+scale\_fill\_manual(values=c("#000000","#FF0000", NA, "black") #NA是空心

+scale\_shape\_manual(values=c(21，24) #21 24号shape

回归线

+ stat\_smooth(method = lm)

默认method是曲线拟合

+ stat\_smooth(method = lm, se=FALSE) #线旁边没有95%CI

加标注：

+annotate("text",label="r^2==0.4",parse=TRUE, x=16, y=52) #==和parse只是为了写出上标小2，xy是那句text在图上的位置

# R PLOTMATH 写出好看的数学标记

+annotate("text",label="AAA", x=16, y=52) #在散点图的一个点标注他的信息

也可以用+geom\_text(aes(label=要标注的变量, x=横坐标变量+1, y=纵坐标变量+1), size=4) #每个点都加上标注的变量，xy调整text所在位置

散点图矩阵（不用ggplot）

plot(dataname[,2:5]) #第2-5列变量作两两交叉的散点图矩阵

百分比堆积图

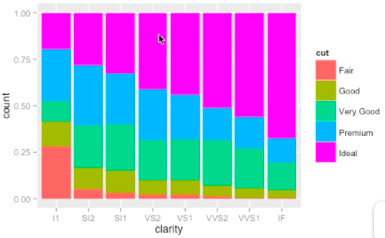
ggplot(liverdata, aes(fat,fill=sex))+geom\_bar(position="fill")

aes里第一个参数是X，可以是factor or not，fill=必须是factor

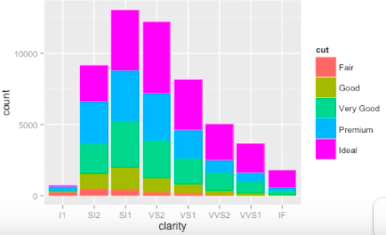
但如果不是factor，会出现空柱子

ggplot(liverdata,aes(factor(glucose.type),fill=sex))+geom\_bar(position="fill")

fill=sex即用sex二色填充, y无意义，position=fill为堆积百分图

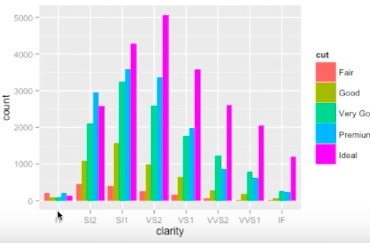


position="stack" 堆积百分图但不是一样高的堆积。



ggplot(liverdata,aes(x=factor(glucose.type),y=fpg,fill=sex))+geom\_bar(position="dodge",stat="identity")

若设定y，则相当于作grouping柱状图，position="dodge" 为并排的bar



当y为连续变量时，必须设定stat="identity"，y为分类变量则不用

层+层

p=p+geom\_point() #点图

可以封装回归线的参数

a=geom\_smooth(method = lm，se=T,color="blue",alpha=0.5,size=2) #需要用时加上

颜色修饰层

ggplot(liverdata,aes(x=factor(glucose.type),y=fpg,fill=sex))+geom\_bar(position="dodge",stat="identity")+scale\_fill\_manual(values=c("#000000","#FF0000"),guide=FALSE) 手动改颜色 guide图例有无

ggplot()+geom\_boxplot() 做箱型图

geom\_bar(width=0.5, position=position\_dodge(0.7)) #前面 单个bar的间距，后面 同group的bar的间距

bar上标数字：

+geom\_text(aes(label= ?), vjust=1.5, colour="white") # vjust 填正值为text写在bar内，负值为bar外

调整y高度

+ylim(0, ???)

# R graphic cookbook 推荐教材

---------------------------------------------------------------------------------------------------

变量转换（后面加一列新变量）

liverdata<-transform(liverdata,v=age^2)

筛选亚组

liverdata2=subset(liverdata,sex==1&bmi>30)

或者subset（dataframe，select=-V2） #去掉V2整列

变量转换类型

asfactor<-function(a,b){ #transfer v as factor

num<-which(colnames(a)==b)

#print(class(a))

#print(num)

a[,num]<-as.factor(as.character(a[,num]))

#print(class(a[,num]))

return(a)

}

asnum<-function(a,b){ #transfer v as numeric

num<-which(colnames(a)==b)

a[,num]<-as.numeric(a[,num])

return(a)

}

listf<-c("sex","glucosetype","glucosetype123","age50","met","bmi28","astalt0.8","nas4","progfibro","nasgroup","safgroup","dm02","antidiabetic",)

# add need to be factor

for (i in c(1:length(listf))){

liverdata2<-asfactor(liverdata2,listf[i])

}

listn<-c("bmr","dmperiod","obesitydegree","flhis","ppg120","apoa","rimt","tpoab","tgab", # add need to be num

"fat","necro","ballooning","nas","saf","uwbc","e2")

for (j in c(1:length(listn))){

liverdata2<-asnum(liverdata2,listn[j])

}

max等函数中加入 na.rm=TRUE 就是不考虑na

否则NA会被当做最大值

if else的写法：

a=3

b=0

if(a>10){

b=1

}else if(a<8){

b=2

}else{

b=3

}

print(b)

a=10

b=0

if(a>10){

b=1}else{

b=2

}

print(b)

mac csv读取

library(foreign, pos=4)

getwd() #

workL <- "/Users/sunxy/Desktop"

setwd(workL)

data2 <- read.csv("bmd.csv", header = T, stringsAsFactors = F,fileEncoding = 'GBK') #mac虚拟机过来的csv

粗略看table 1

library(table1)

table1(~age +bmi+fpg+tc+tg+hdl+ldl+alt+ast+ggt + nfs | group, data=data3)

自己写的table1

write\_v\_table1<-function(vname,n,groupcount,data){

library(agricolae)

i<-which(colnames(data)==vname)

model <- aov(data[,i] ~ group, data)

p.anova<-summary(model)[[1]][1,5] # anova p

out <- LSD.test(model, "group", p.adj = "bonferroni" ) # multiple comparision

#out$groups$groups #group difference a b c ab

v=c(vname)

for (gn in c(1:groupcount)){

if (n==1){

v<-append(v,paste(round(out$means[gn,1],2),"±",round(out$means[gn,2],2)))

}else{

v<-append(v,paste(round(out$means[gn,9],2),"(",round(out$means[gn,8],2),"-",round(out$means[gn,10],2),")"))

}

v<-append(v,paste(out$groups$groups[gn])) #group difference a b c ab

}

v<-append(v,round(p.anova,3)) # anova p

v<-append(v,write\_meansd(vname,n,data)) #total count

return(v)

}

# example

# write\_v\_table1("age",2,3,data3) #vname, #1=mean sd,2=(25~75) #how many groups #dataframe

# rbind(write\_v\_table1("bmi2",1,3,data3),write\_v\_table1("age",1,3,data3))

相关矩阵可视化筛选

library(PerformanceAnalytics)

chart.Correlation(data3[,c(81,55:57,63,66,64,65,67)], histogram=TRUE, pch=19)

获得相关系数及显著性

library(psych)

cor\_matrix<-data3[,c(81,55,83,84,85,87,98,92,91,93:97)]

# corr.test(cor\_matrix)

# corr.test(cor\_matrix)$r[,1] # [,1]first column

# corr.test(cor\_matrix)$p[,1]

# names(corr.test(cor\_matrix)$p[,1])

cor\_result<-rbind(names(corr.test(cor\_matrix)$p[,1]),round(corr.test(cor\_matrix)$r[,1],3),round(corr.test(cor\_matrix)$p[,1],3))

cor\_result

获得偏相关系数及显著性

library(ggm)

library(psych)

mypcor<-function(a){

v<- which(colnames(data3)==a)

test=data3[,c(81,v,74,76,77)] # former 2 for correlation, later for adjust

r<-pcor(c(1,2,3,4,5),cov(test,use = "na.or.complete")) #NA exclude！！！must cov里写method="spearman"

print(r)

p.pcor<-pcor.test(r,3,236)[3] #3=num of adjust v #236=sample size.

print(p.pcor)

}

mypcor("pp")

去除最大值

#extreme value

max(data3$ost2,na.rm=TRUE)

which(data3$ost2==63.5)

data3$ost2[227]<-NA

散点图垂直翻转

#nfs no change, Flip vertically pth

nfsrank<-rank(data3$nfs2)

#a<-which(colnames(data3)=="pth")

data3$ost2<-data3$ost

for (i in c(1:236)){

print(i)

b=which(nfsrank==(237-nfsrank[i]))[1]

print(b)

if (is.na(data3$ost[i])==FALSE & is.na(data3$ost[b])==FALSE){

data3$ost2[i]<-data3$ost[b]

print(i)

print(data3$ost2[i])

}

}

qplot(data3$nfs2,data3$ost2)+ geom\_smooth(method = lm)

一键获得多组的均值 anova及两两比较

#in dataframe,group="group"

write\_v\_table1<-function(vname,n,groupcount,data){

library(agricolae)

i<-which(colnames(data)==vname)

model <- aov(data[,i] ~ group, data)

p.anova<-summary(model)[[1]][1,5] # anova p

out <- LSD.test(model, "group", p.adj = "bonferroni" ) # multiple comparision

#out$groups$groups #group difference a b c ab

v=c(vname)

for (gn in c(1:groupcount)){

if (n==1){

v<-append(v,paste(round(out$means[gn,1],2),"±",round(out$means[gn,2],2)))

}else{

v<-append(v,paste(round(out$means[gn,9],2),"(",round(out$means[gn,8],2),"-",round(out$means[gn,10],2),")"))

}

v<-append(v,paste(out$groups$groups[gn]))

}

v<-append(v,p.anova)

return(v)

}

# example

# write\_v\_table1("age",2,3,data3) #vname,"" #1=mean sd,2=(25~75) #how many groups #dataframe

# rbind(write\_v\_table1("bmi2",1,3,data3),write\_v\_table1("age",1,3,data3))

读spss

library(foreign)

mydata<-read.spss('全国调研加地区-0824xmf\_1-0107加烟酒.sav')

读excel

library(readxl)

#data <- read.table("name.txt",header = T,sep = "")

data <- read\_excel("全国调研 给潇泱.xlsx",sheet=1,col\_names = T,col\_types = NULL ,na="", skip=0)

data[823,1]<-"D03-088-2" #编号重复

matrix写入excel

library(xlsx)

write.xlsx(table1,"table1.xlsx")

merge 数据集的合并

mydata3 <- merge(mydata2,tatin\_data,by.tatin\_data = "编号",by.mydata2 = "编号",all.x=TRUE) #前一个小名单，后一个大名单，all.x=true 保留第一个矩阵的结构

查找字符串中是否含有某字符串

grep("他汀",colnames(data)) 显示几个数字

交叉表

my\_crosstab<-function(dataname,vname,groupname){

crosstab<-table(dataname[,c(vname,groupname)]) #分组变量写后面

crosstab<-as.matrix(crosstab)

chi2\_p<-round(chisq.test(crosstab)$p.value,3)

l=length(crosstab)

crosstab2<-crosstab

for (n in c(1:l)){

#print(n)

x<-crosstab[n]

colnum<-which(crosstab==x,arr.ind = TRUE)[2]#获取元素x在表中的行列号 [2]为在第几列

#print(colnum)

prop<-x/apply(crosstab,2,sum)[colnum]

#print(prop)

y=paste(x,"(",round(prop\*100,2), "%",")",sep='')

#print(y)

crosstab2[n]<-y

#print(y)

}

#print(crosstab2)

return(crosstab2)

#return(chi2\_p)

#print(chi2\_p)

} # NXN 二级列联表及

my\_crosstab\_p<-function(dataname,vname,groupname){

crosstab<-table(dataname[,c(vname,groupname)]) #分组变量写后面

chi2\_p<-round(chisq.test(crosstab)$p.value,3)

return(chi2\_p)

}

正常交叉表：

b<-table(Arthritis$Sex,Arthritis$Improved)

addmargins(b) #添加横竖sum

ggplot具体图的设置

http://www.cookbook-r.com/Graphs/Legends\_%28ggplot2%29/

https://blog.csdn.net/skyonefly/article/details/50708690?depth\_1-utm\_source=distribute.pc\_relevant.none-task&utm\_source=distribute.pc\_relevant.none-task

https://blog.csdn.net/zx403413599/article/details/48581713?depth\_1-utm\_source=distribute.pc\_relevant.none-task&utm\_source=distribute.pc\_relevant.none-task

非ggplot可以设置分离坐标轴

http://blog.sina.com.cn/s/blog\_4c7fada80102x3gd.html

dplyr 数据处理，主要是代码更简洁