Handout 4. Descriptive statistics and graphics

1. Summary statistics for a single group

It is easy to calculate simple summary statistics with R. Here is how to calculate the mean, standard deviation, variance, and median.

```
> x <- rnorm(50)
> mean(x)
> sd(x)
> var(x)
sd(x)^2
> median(x)
> quantile(x)
> pvec <- seq(0,1,0.1)
> quantile(x,pvec)
 quantile(x,c(0.1,0.4))
```

```
library(ISwR)
data(juul)
juul
dim(juul)
head(juul)
juul[1:10,]

attach(juul)
names(juul)

mean(igf1)
igf1
mean(igf1,na.rm=T)
sum(igf1)

sum(igf1,na.rm=T)
sum(is.na(igf1))
sum(!is.na(igf1))

summary(igf1)
summary(juul)
```

```
juul<-transform(juul,sex=factor(sex,
labels=c("f","m")),menarche=factor(menarche,
labels=c("no","yes")),tanner=factor(tanner,
labels=c("one","two","three","four","five")))

juul$menarche
```

2. Graphic display of distributions --- Histograms, empirical distributions, Q-Q plot, Boxplot

```
x<-rnorm(50)

hist(x)
```

```
x<-c(1.5,2.5,3.5,4.5,5.5,6.5,8.5,9.5,12.5)
y<-c(5,7,12,2,1,4,14,2,3)
z<-rep(x,y)
brk<-c(0,1,2,3,5,7,9,10,11,13)
hist(z,breaks=brk)
```

```
> ### empirical distribution function
> x<-rnorm(100)
> n <- length(x)
>   plot(sort(x),(1:n)/n,type="s",ylim=c(0,1))
plot(sort(x),(1:n)/n,type="l",ylim=c(0,1))
#plot(x,pnorm(x))

empirical cumulative distribution function is defined as the
fraction of data smaller than or equal to x.
That is , if x is the kth smallest observation, the the
proportion k/n of the data is smaller than or equal to x



#qq plot
qqnorm(x)
```

BoxPlot

```
data(IgM)
par(mfrow=c(1,2))
boxplot(IgM)
boxplot(log(IgM))
par(mfrow=c(1,1))
```

mfrow graphical parameter should read as" multiframe, rowwise, 1*2 layout" individual plots are organized in 1 row and 2 columns.

```
par(mfrow=c(1,2))
boxplot(log(IgM))
boxplot(IgM)
```

```
par(mfcol=c(2,2))
boxplot(log(IgM))
boxplot(IgM)
boxplot(sin(IgM))
boxplot(cos(IgM))
```

```
par(mfrow=c(2,2))
boxplot(log(IgM))
boxplot(IgM)
boxplot(sin(IgM))
boxplot(cos(IgM))
```

3. Summary statistics by groups

```
data(red.cell.folate)
attach(red.cell.folate)
xbar=tapply(folate,ventilation,mean)
s=tapply(folate,ventilation,sd)
n=tapply(folate,ventilation,length)
cbind(mean=xbar,std.dev=s,n=n)
```

```
data(juul)
tapply(igf1,tanner,mean)
tapply(igf1,tanner,mean,na.rm=T)
```

na.rm=t as a parameter to mean to make it exclude the missing values

4. Graphics for grouped data

Histograms

```
  data(energy)
  attach(energy)
 expend.lean<-expend[stature=="lean"]
   expend.obese<-expend[stature=="obese"]
par(mfrow=c(2,1))
hist(expend.lean,breaks=10,xlim=c(5,13),ylim=c(0,4),col="white")
hist(expend.obese,breaks=10,xlim=c(5,13),ylim=c(0,4),col="grey")
par(mfrow=c(1,1))
```

Parallel boxplot

```
boxplot(expend~stature)
boxplot(expend.lean , expend.obese)
```

Stripcharts

```
opar<-par(mfrow=c(2,2),mex=0.8,mar=c(3,3,2,1)+.1)
stripchart(expend~stature)
stripchart(expend~stature,method="stack")
stripchart(expend~stature,method="jitter")
stripchart(expend~stature,method="stack",jitter=0.03)
par(opar)  #reestablished
```

5. Tables

Categorical data are usually described in the form of tables. A two-way table can be entered as a matrix object.

```
> caff.marital <- matrix(c(652,1537,598,242,36,46,38,21,218,
327,106,67), nrow=3,byrow=T)
> colnames(caff.marital) <- c("0","1-150","151-300",">300")
> rownames(caff.marital) <-
c("Married","Prev.married","Single")
> caff.marital
                 0    1-150 151-300 >300
Married          652  1537  598     242
Prev.married     36   46    38      21
Single           218  327   106     67
```

Furthermore, you can name the row and column names as follows. This is particularly useful if you are generating many tables with similar classification criteria.

```
> names(dimnames(caff.marital)) <- c("marital","consumption")
> caff.marital
             consumption
marital           0    1-150 151-300 >300
Married           652  1537  598     242
Prev.married      36   46    38      21
Single            218  327   106     67
```

Like any matrix, a table can be transposed with the t function:

```
> t(caff.marital)
```

Exercise: Construct the following table which summarize the number of people smoking and nonsmoking in a class.

	Smoking	Nonsmoking
Male	23	45
Female	34	54

```
data(juul)
attach(juul)
table(sex)
table(sex,menarche)
table(menarche,tanner)

table(sex,menarche,tanner)

table(tanner,sex)
margin.table(table(tanner,sex),1)
margin.table(table(tanner,sex),2)

prop.table(table(tanner,sex),1)
prop.table(table(tanner,sex),2)

table(tanner,sex)/sum(table(tanner,sex))  #grand total of the
table
```

6. Graphical display of tables

6.1 barplot

```
barplot(prop.table(t(caff.marital)),legend.text=colnames(caff.marital),col=c("white","blue","green","black
"))
```

6.2 dotcharts

```
dotchart(t(caff.marital))
```

6.3 pie charts

```
opar<- par (mfrow=c(2,2), mex=0.8, mar=c(1,1,2,1))
slices<- c("white", "grey80", "grey50", "black")
pie(caff.marital["Married",], main="Married", col=slices)
pie(caff.marital["Prev.married",],   main="Previously married", col=slices)
pie(caff.marital["Single",], main="Single", col=slices)
par(opar)
```