Friedrich-Alexander-Universität
Erlangen-Nürnberg

Lehrstuhl für Statistik und Ökonometrie

Diskussionspapier
83 / 2008

Some R graphics for bivariate distributions

Ingo Klein

Lange Gasse 20 · D-90403 Nürnberg

Some R graphics for bivariate distributions

Ingo Klein

Lehrstuhl für Statistik und Ökonometrie
Universität Erlangen-Nürnberg
ingo.klein@wiso.uni-erlangen.de

Abstract

There is no package in R to plot bivariate distributions for discrete variables or variables given by classes. Therefore, with the help of the already implemented R routine "persp" R functions will be proposed for 3-D plots of the bivariate distribution of discrete variables, the so-called stereogram that generalizes the well-known histogram for cross-classified data and the approximative bivariate distribution function for cross-classified data.

Keywords: Bivariate distribution function, bivariate frequency density, stereogram, approximate bivariate distribution function

1 Introduction

There is no package in R to plot bivariate distributions for discrete variables or variables given by classes. Therefore, with the help of the R routine "persp" R functions will be proposed for 3-D plots of the bivariate distribution of discrete variables, the so-called stereogram that generalizes the well-known histogram for cross-classified data and the approximative bivariate distribution function for cross-classified data. The paper is organized in the following way. First we introduce the notation for bivariate distribution functions of discrete variables. Then the R programming code will be presented. Afterwards an example will be given that demonstrate how the R graphics can be applied. This procedure will repeated for the stereogram and the approximated bivariate distribution function for cross-classified data.

2 Bivariate distribution function for discrete variables

2.1 Two kinds of data

The data can occur in two different ways. In the first case we already know the frequency (or probability) table, the bivariate distribution can be calculated from directly. In a second case we observe the original data points. From these the frequency table has to be calculated before in a second step the bivariate distribution function can be calculated in the same way as in the first case.

Case 1: Let us consider a population of size n and two quantitative variables X and Y. The possible values of X (Y) are $x_1 < x_2 < \ldots, x_k$ ($y_1 < y_2 < \ldots, y_l$), $k, l \in \mathbb{N}$. f_{ij} is the relative frequency the pair of possible values (x_i, y_j), $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$ will be observed in the population. The bivariate cumulative distribution function of (X, Y) is defined by

$$F_{X,Y}(x, y) = \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} I[x_i \leq x, y_j \leq y] \;\; (x, y) \in \mathbb{R}^2.$$

$I[.,.]$ denotes the indicator function taking the value 1 if the expression in parentheses is true and 0 otherwise. For the possible values (x_i, y_j) the values

of bivariate cumulative distribution function F_{ij} are given by the recurrence relation

$$F_{ij} = f_{ij} + F_{i,j-1} + F_{i-1,j} - F_{i-1,j-1}\ ; i = 1, \ldots, k,\ j = 1, 3, \ldots, l$$

with $F_{0,j} = F_{i,0} = 0$ for $i = 0, 1, \ldots, k,\ j = 0, 1, \ldots, l$.

Case 2: If the elements of the population are $e_1, e_2, \ldots, e_n$ we can observe the data points $(X(e_\nu), Y(e_\nu))$, $\nu = 1, 2, \ldots, n$. In a first step from these data points the possible values $x_1 < x_2 < \ldots < x_k$ of X and $y_1 < y_2 < \ldots < y_l$ of Y, $1 \leq k, l \leq n$ can be calculated. n_{ij} (f_{ij}) is the absolute (relative) frequency the pair (x_i, y_j) occurs in $\{(X(e_\nu), Y(e_\nu)) | \nu = 1, 2, \ldots, n\}$ for $i = 1, 2, \ldots, k,\ j = 1, 2, \ldots, l$. In R we can use the commands factors to extract the possible different values of the two variables X and Y and table to calculate the frequency table.

2.2 R-program cdf2

```
function(xx,yy,f,xachse,yachse)
{
# 3-D plot of the bivariate distribution function
# for pairs of observed data or for a table with
# absolute or relative frequencies (probabilities)
# for two discrete variables
#
# function name: plot.cdf2
# Plotting the bivariate distribution function for
# two discrete variables
# if paires of observed data (case 2) or a frequency
# (probability) table (case 1) are given
#
# Arguments:
# xx: observed data or categories of the first
#     discrete variable
# yy: observed data or categories of the second
#     discrete variable
# f : Case 1: f=0 (especially length(f)=1) means
#     observed data
#   : Case 2: length(f)>1 means frequency (probability)
```

```
#   :         table
# xachse: x-axis name
# yachse: y-axis name
#
if (length(f)>1)
{
# Case 1: Frequency (probability) table
    xi=sort(xx)
    yj=sort(yy)
}
else
{
# Case 2: Observed data from which the frequency
# table has to be calculated first
    xi=as.numeric(levels(as.factor(xx)))
    yj=as.numeric(levels(as.factor(yy)))
    f=table(xx,yy)
}
#
# Calculating relative frequencies
if (sum(sum(f)) > 1){f=f/sum(sum(f))}
F=f
# Bivariate distribution function values for the
# categories of the discrete variables
F[1,]=cumsum(f[1,])
F[,1]=cumsum(f[,1])
k=length(xi)
l=length(yj)
for (i in 2:k){
     for (j in 2:l){
          F[i,j]=f[i,j]+F[i-1,j]+F[i,j-1]-F[i-1,j-1]
                   }
              }
#
#
# Bivariate distribution function values for all
# arguments (x,y) needed for the 3-D plot
deltax=(max(xi)-min(xi))/200
```

```
deltay=(max(yj)-min(yj))/200
x=seq(min(xi)-deltax,max(xi)+deltax,deltax)
y=seq(min(yj)-deltay,max(yj)+deltay,deltay)
n1=length(x)
n2=length(y)
z=matrix(rep(0,n1*n2),ncol=n2)
for (i in 1:n1){
    for (j in 1:n2){
         i1=(x[i]>=xi)
         i2=(y[j]>=yj)
         if (sum(i1) == 0 | sum(i2)==0){z[i,j]=0}
         if (sum(i1) >= k & sum(i2) >=l){z[i,j]=1}
         if (sum(i1) >= k & sum(i2) < l & sum(i2) > 0){z[i,j]=F[k,sum(i2)]}
         if (sum(i1) < k  & sum(i2) >=l & sum(i1) > 0){z[i,j]=F[sum(i1),l]}
         if (sum(i1) < k  & sum(i2) < l & sum(i1) > 0 & sum(i2) > 0)
+         {z[i,j]=F[sum(i1),sum(i2)]}
                    }
                }
# 3-D plot
persp(x,y,z,theta=-30,phi=15,col="red",shade=0.1,main="Bivariate distribution
+ function \n for discrete variables",xlab=xachse,ylab=yachse,zlab="",
+ cex.axis=0.75,ticktype="detailed")
}
```

2.3 Example 1

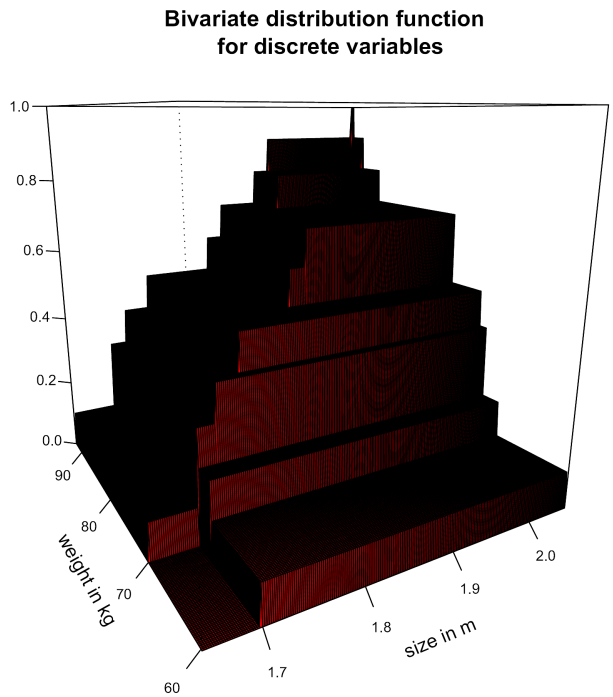
We measure size X (in cm) and weight Y (in kg) of $n = 10$ participants of a basic course in statistics. We get the following data

e_ν	1	2	3	4	5	6	7	8	9	10
$X(e_\nu)$	1.83	1.72	1.65	1.70	2.05	1.92	1.85	1.70	1.75	1.90
$Y(e_\nu$	75	70	70	60	90	92	75	68	71	87

The followings commands

```
> x=c(1.83,1.72,1.65,1.70,2.05,1.92,1.85,1.70,1.75,1.9)
> y=c(75,70,70,60,90,92,75,68,71,87)
> plot.cdf2(x,y,f=0,"size in m","weight in kg")
```

give



2.4 Example 2

Now, we throw a green and a red die. Let G and R be the spot of the green and the red die. We consider the random variables $X = G + R$ (sum of the spots) and $Y = \min\{G, R\}$ (smaller spot). The possible values for X (Y) are $x_1 = 2, x_2 = 3, \ldots, x_{11} = 12$ ($y - 1 = 1, y_2 = 2, \ldots, y_6 = 6$). The table of probabilities $p_{ij} =$

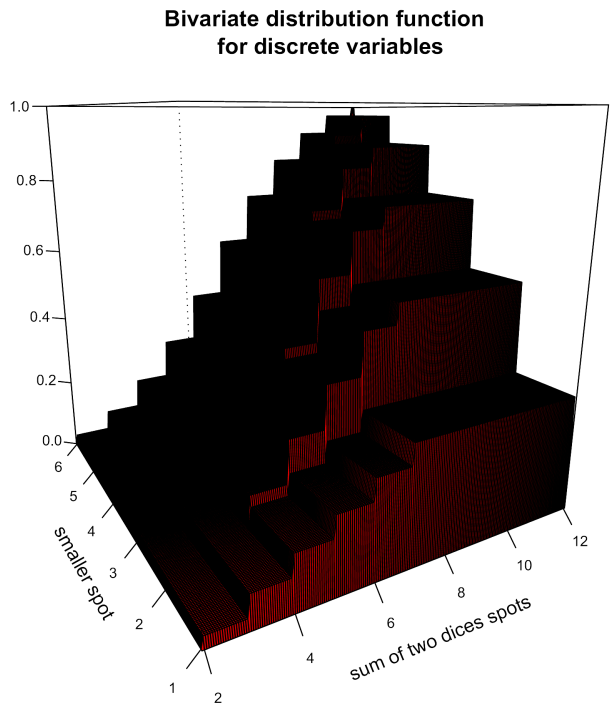
$P(X = x : i, Y = y_j)$, $i = 1, 2, \ldots, 11$, $j = 1, 2, \ldots, 6$ is been given by

$X \backslash Y$	1	2	3	4	5	6	$\sum$
2	1/36	0	0	0	0	0	1/36
3	2/36	0	0	0	0	0	2/36
4	2/36	1/36	0	0	0	0	3/36
5	2/36	2/36	0	0	0	0	4/36
6	2/36	2/36	1/36	0	0	0	5/36
7	2/36	2/36	2/36	0	0	0	6/36
8	0	2/36	2/36	1/36	0	0	5/36
9	0	0	2/36	2/36	0	0	4/36
10	0	0	0	2/36	1/36	0	3/36
11	0	0	0	0	2/36	0	2/36
12	0	0	0	0	0	1/36	1/36
$\sum$	11/36	9/36	7/36	5/36	3/36	1/36	1

The commands

```
>x=seq(2,12)
>y=seq(1,6)
>p=c(1/36,0,0,0,0,0,
 2/36,0,0,0,0,0,
 2/36,1/36,0,0,0,0,
 2/36,2/36,0,0,0,0,
 2/36,2/36,1/36,0,0,0,
 2/36,2/36,2/36,0,0,0,
 0,2/36,2/36,1/36,0,0,
 0,0,2/36,2/36,0,0,
 0,0,0,2/36,1/36,0,
 0,0,0,0,2/36,0,
 0,0,0,0,0,1/36)
>p=matrix(F,byrow=TRUE,ncol=6)
>plot.cdf2(x,y,p,"sum of two dices spots","smaller spot")
```

give the following figure for the bivariate distribution function of the two discrete random variables X and Y:



3 Stereogram and the approximate bivariate distribution function of data given in classes

3.1 Preliminaries

Let X and Y be two variables for which we can only observe that the data points $(X(e_\nu), Y(e_\nu))$, $\nu = 1, 2, \ldots, n$ fall in some class or not. The breakpoint for the k (l) classes of X (Y) are given by $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{k-1}$ ($\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{l-1}$) for $k, l > 1$. With $\tilde{x}_0$ ($\tilde{y}_0$) we denote a lower limit of the first class and with $\tilde{x}_k$ ($\tilde{y}_l$) the upper limit of the last class. The width of the classes are given by $\Delta x_i = \tilde{x}_i - \tilde{x}_{i-1}$, $i = 1, 2, \ldots, k$ and $\Delta y_j = \tilde{y}_j - \tilde{y}_{j-1}$, $j = 1, 2, ldots, l$.

n_{ij} is the number of data points $(X(e_\nu), Y(e_\nu))$ falling in class i of X and class j of Y. $f_{ij} = n_{ij}/n$ is the corresponding relative frequency for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, l$.

The following so-called correlation table describes the structures of the data:

$X \setminus Y$	$(\tilde{y}_0, \tilde{y}_1]$	$(\tilde{y}_1, \tilde{y}_2]$	$\dots$	$(\tilde{y}_{l-1}, \tilde{y}_l]$	
$(\tilde{x}_0, \tilde{x}_1]$	f_{11}	f_{12}	$\dots$	f_{1l}	$f_{1\cdot}$
$(\tilde{y}_1, \tilde{x}_2]$	f_{21}	f_{22}	$\dots$	f_{2l}	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(\tilde{x}_{k-1}, \tilde{x}_k]$	f_{k1}	f_{k2}	$\dots$	f_{kl}	$f_{k\cdot}$
$\sum$	$f_{\cdot 1}$	$f_{\cdot 2}$	$\dots$	$f_{\cdot l}$	1

3.2 Stereogram

The so-called stereogram generalizes the well-known histogram (see f.e. Wetzel (1971, p. 58). For one variable X the histogram draws a plot of the coordinates $(x, f_X^*(x))$, $x \in \mathbb{R}$ with

$$f_X^*(x) = \frac{f_i}{(\tilde{x}_{i\cdot} - \tilde{x}_{i-1})} \quad \text{for } \tilde{x}_{i-1} \leq x < \tilde{x}_i.$$

f^* has the properties of a density and will be called the "frequency density". Especially, the frequency density gives no frequencies, but the areas under the frequency density are again frequencies (see f.e. Dalgaard (2002), p. 61, Vogel (2000), p. 18).

The bivariate frequency density of two variables X and Y with in $k \times l$ classes is defined by

$$f_{X,Y}^*(x, y) = \frac{f_{ij}}{\Delta x_i \Delta y_j} \quad \text{for } \tilde{x}_{i-1} \leq x < \tilde{x}_i,\ \tilde{y}_{j-1} \leq y < \tilde{y}_j$$

Now, the stereogram is given by the plot of $(x, y, f_{X,Y}^*(x, y))$, $x, y \in \mathbb{R}$. $f_{X,Y}^*$ Volumes under $f_{X,Y}^*$ are again frequencies.

3.3 R-program stereo

```
function(kgx,kgy,f,xachse,yachse)
{
# 3-D plot of the bivariate frequency function
# (=stereogram)
# function name: stereo
#
# Arguments:
# kgx: breakpoints of the first variable (incl. lower and upper
#      limit)
```

```
# kgy: breakpoints of the second variable (incl. lower and upper
#      limit)
# f  : absolute or relative frequencies for the k times l classes
# xachse: x-axis name
# yachse: y-axis name
#
# Calculating relative frequencies
f=f/sum(sum(f))
# Calculating the number of classes from the number of breakpoints
k=length(kgx)-1
l=length(kgy)-1
# Calculating the width of the classes
kbx=diff(kgx)
kby=diff(kgy)
kb=kbx%*%t(kby)
# Calculating the values of the frequency density for all classes
fd=f/kb
# Generating the arguments the stereogramm will be plotted for
deltax=(max(kgx)-min(kgx))/300
deltay=(max(kgy)-min(kgy))/300
x=seq(min(kgx)-deltax,max(kgx)+deltax,deltax)
y=seq(min(kgy)-deltay,max(kgy)+deltay,deltay)
nx=length(x)
ny=length(y)
# Calculating the frequency density for all arguments
fdxy=matrix(rep(0,nx*ny),byrow=TRUE,ncol=ny)
for (i in 1:nx){
    for (j in 1:ny){
         ix=(x[i]<=kgx)
         kx=k+1-sum(ix)
         jy=(y[j]<=kgy)
         ly=l+1-sum(jy)
         fdxy[i,j]=0
         if (kx>0 & ly>0 & kx < k+1 & ly < l+1)
            {
            fdxy[i,j]=fd[kx,ly]
            }
                   }
```

```
                }
# Plotting the stereogram
ueber=paste("Stereogram for two variables \n with k=",k"," times l=",l," classes")
persp(x,y,fdxy,zlim=c(0,max(fdxy)+0.05),theta=-30,phi=15,col="red",shade=0.5,
+ticktype="detailed",main=ueber,xlab=xachse,ylab=yachse,zlab="")
}
```

3.4 Example 3

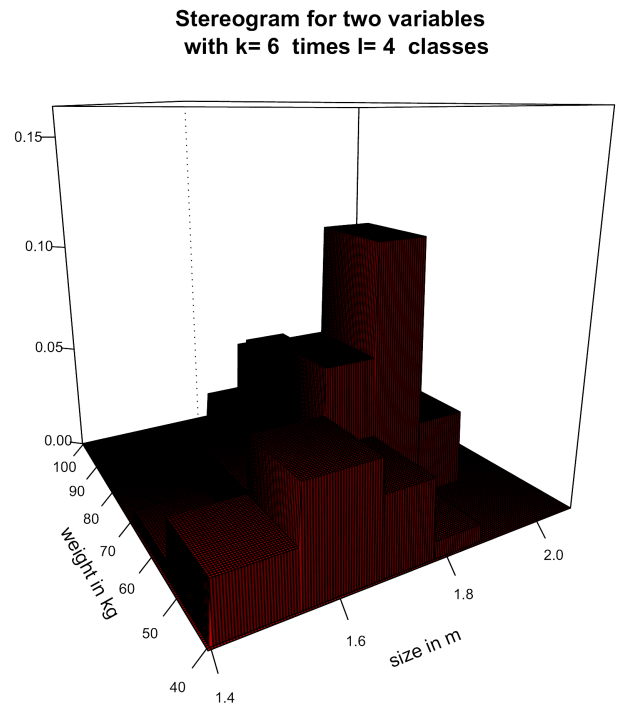
Let X and Y be again the size (in m) and the weight (in kg) of $n = 200$ students. The data is given by 5×4 classes with the following relative frequencies (correlation table):

	j	1	2	3	4	
	$(\tilde{y}_{j-1}; \tilde{y}_j]$	$(40; 55]$	$(55; 70]$	$(70; 85]$	$(85; 100]$	
i	$(\tilde{x}_{i-1}; \tilde{x}_i]$					$f_{i\cdot}$
1	$(1.40; 1.54]$	0.05	0.01	0	0	0.06
2	$(1,54; 1,68]$	0.10	0.05	0.04	0	0.19
3	$(1.68; 1.78]$	0.05	0.10	0.08	0.02	0.25
4	$(1.78; 1.88]$	0.01	0.17	0.05	0.06	0.29
5	$(1.88; 1.98]$	0	0.05	0.08	0.02	0.15
6	$(1.98; 2.08]$	0	0	0.02	0.04	0.06
	$f_{\cdot j}$	0.21	0.38	0.27	0.14	1

The following commands

```
clx=c(1.40,1.54,1.68,1.78,1.88,1.98,2.08)
cly=c(40,55,70,85,100)
f=matrix(c(0.05,0.01,0,0,0.10,0.05,0.04,0,0.05,0.10,0.08,0.02,
+ 0.01,0.17,0.05,0.06,0,0.05,0.08,0.02,0,0,0.02,0.04),ncol=4,byrow=TRUE)
stereo(clx,cly,f,"size in m","weight in kg")
```

result in the stereogram



3.5 Approximative bivariate distribution function

Denote the cumulated relative class frequencies with

$$F_{ij} = \sum_{u=1}^{i} \sum_{v=1}^{j} f_{uv}, \quad i = 1, 2, \ldots, k, \; j = 1, 2, \ldots, l.$$

The approximative bivariate distribution function $F^*_{X,Y}$ of the variables X and Y is given by

$$F^*_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f^*_{X,Y}(u, v) dv du.$$

This is the volume under the stereogram until (x, y).

Denote the cumulated relative class frequencies with

$$F_{ij} = \sum_{u=1}^{i} \sum_{v=1}^{j} f_{uv}, \quad i = 1, 2, \ldots, k, \; j = 1, 2, \ldots, l.$$

Substituting the formula for the bivariate frequency curve and integrating leads to

$$F^*_{X,Y}(x,y) = F_{i-1,j-1} + (F_{i,j-1} - F_{i-1,j-1})\frac{x-\tilde{x}_{i-1}}{\Delta x_i} + (F_{i-1,j} - F_{i-1,j-1})\frac{y-\tilde{y}_{j-1}}{\Delta y_j}$$
$$+ f_{ij}\frac{(x-\tilde{x}_{i-1})(y-\tilde{y}_{j-1})}{\Delta x_i \Delta y_j}$$

if $\tilde{x}_{i-1} \leq x < \tilde{x}_i$ and $\tilde{y}_{j-1} \leq y < \tilde{y}_j$, $i = 1, 2, \ldots, k$. For $x < \tilde{x}_0$ or $y < \tilde{y}_0$ we set $F^*_{X,Y}(x,y) = 0$. $F^*_{X,Y}(x,y) = 1$ for $x > \tilde{x}_k$ and $y > \tilde{y}_l$. The remaining cases are

$$F^*_{X,Y}(x,y) = F_{i-1,l} + f_{i\cdot}\frac{x-\tilde{x}_{i-1}}{\Delta x_i} \quad \text{for } \tilde{x}_{i-1} \leq x < \tilde{x}_i,\ y > \tilde{y}_l$$

for $i = 1, 2, \ldots, k$ and

$$F^*_{X,Y}(x,y) = F_{k,j-1} + f_{\cdot j}\frac{y-\tilde{y}_{j-1}}{\Delta y_j} \quad \text{for } x > \tilde{x}_k,\ \tilde{y}_{j-1} \leq y < \tilde{y}_j$$

for $j = 1, 2, \ldots, l$.

3.6 R program: plot.approxcdf2

```
function(kgx,kgy,f,xachse,yachse)
{
# 3-D plot of the bivariate frequency function
# (=stereogram)
# function name: plot.approxcdf2
#
# Arguments:
# kgx: breakpoints of the first variable (incl. lower and upper
#      limit)
# kgy: breakpoints of the second variable (incl. lower and upper
#      limit)
# f  : absolute or relative frequencies for the k times l classes
# xachse: x-axis name
# yachse: y-axis name
#
# Calculating relative frequencies
f=f/sum(sum(f))
#
# Calculating the number of classes from the number of breakpoints
```

```
k=length(kgx)-1
l=length(kgy)-1
fx=rep(0,k)
fy=rep(0,l)
for (i in (1:k)){fx[i]=sum(f[i,])}
for (j in (1:l)){fy[j]=sum(f[,j])}
# Calculating the width of the classes
kbx=diff(kgx)
kby=diff(kgy)
# Calculating the values of the frequency density for all classes
#
# Bivariate distribution function values for the
# categories of the discrete variables
F=matrix(rep(0,(k+1)*(l+1)),byrow=TRUE,ncol=l+1)
for (i in (2:(k+1))){
     for (j in (2:(l+1))){
          F[i,j]=f[i-1,j-1]+F[i-1,j]+F[i,j-1]-F[i-1,j-1]
                  }
              }
#
# Generating the arguments the approximated bivariate distribution
# function will be plotted for
deltax=(max(kgx)-min(kgx))/100
deltay=(max(kgy)-min(kgy))/100
x=seq(min(kgx),max(kgx),deltax)
y=seq(min(kgy),max(kgy),deltay)
nx=length(x)
ny=length(y)
#
# Calculating the approximative bivariate distribution function for all arguments
Fdxy=matrix(rep(0,nx*ny),byrow=TRUE,ncol=ny)
for (i in 1:nx){
    for (j in 1:ny){
         ix=(x[i]>kgx)
         kx=sum(ix)
         jy=(y[j]>kgy)
         ly=sum(jy)
         Fdxy[i,j]=0
```

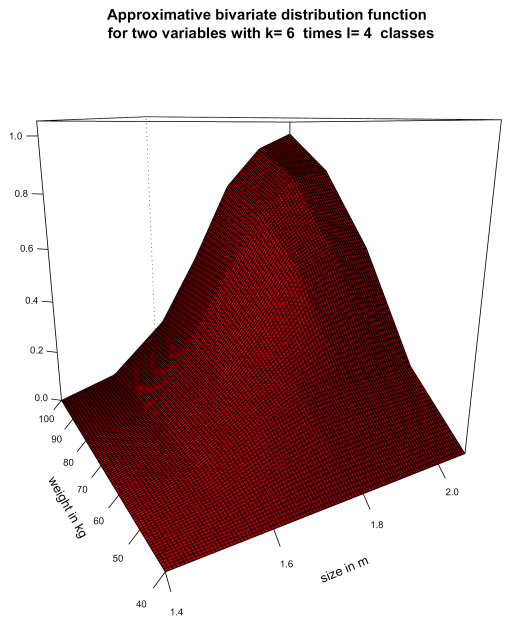
```
         if (kx>0 & ly>0 & kx < k+1 & ly < l+1)
            {
            Fdxy[i,j]=F[kx,ly]+(F[kx+1,ly]-F[kx,ly])*(x[i]-kgx[kx])/kbx[kx]
+            + (F[kx,ly+1]-F[kx,ly])*(y[j]-kgy[ly])/kby[ly]
+            + f[kx,ly]*(x[i]-kgx[kx])*(y[j]-kgy[ly])/(kbx[kx]*kby[ly])
            }
         if (kx>0 & kx < k+1 & ly >= l+1)
            {
            Fdxy[i,j]=F[kx,l+1] + fx[kx]/kbx[kx]*(x[i]-kgx[kx])
            }
         if (kx >= k+1 & ly >0 & ly < l+1)
            {
            Fdxy[i,j]=F[k+1,ly] + fy[ly]/kby[ly]*(y[j]-kgy[ly])
            }
         if (kx >= k+1 & ly >= l+1){Fdxy[i,j]=1}
                     }
               }
#
# Plotting the approximative bivariate distribution function
ueber=paste("Approximative bivariate distribution function \n for two variables
+ with k=",k," times l=",l," classes")
persp(x,y,Fdxy,zlim=c(0,max(Fdxy)+0.05),theta=-30,phi=15,col="red",shade=0.5,
+ticktype="detailed",cex.axis=0.75,main=ueber,xlab=xachse,ylab=yachse,zlab="")
}
```

3.7 Example 4

We continue example 3. The following commands

```
kgx=c(1.40,1.54,1.68,1.78,1.88,1.98,2.08)
kgy=c(40,55,70,85,100)
f=matrix(c(0.05,0.01,0,0,0.10,0.05,0.04,0,0.05,0.10,0.08,0.02,0.01,
+ 0.17,0.05,0.06,0,0.05,0.08,0.02,0,0,0.02,0.04),ncol=4,byrow=TRUE)
plot.approxcdf2(clx,cly,f,"size in m","weight in kg")
```

give the following 3-D plot



4 Summary

There is no R package to plot bivariate distributions for discrete variables or variables given by classes. Therefore, with the help of the already implemented R routine "persp" R functions have been proposed for 3-D plots of the bivariate distribution of discrete variables, the so-called stereogram that generalizes the well-known histogram for cross-classified data and the approximative bivariate distribution function for cross-classified data.

Literature

1. Dalgaard, P. (2002). *Introductory Statistics with R.* Springer, New York,
2. Vogel, F. (2000). *Beschreibende und schließende Statistik.* Oldenbourg, München.
3. Wetzel, W. (1971). *Statistische Grundausbildung für Wirtschaftswissenschaftler. I. Beschreibende Statistik* Walter de Gruyter, Berlin.