

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ



ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Elements of Statistics and Probability

LECTURE 3 – Describing Data

Xanthi Pedeli

Assistant Professor, xpedeli@aueb.gr

Department of Statistics, AUEB

Notes by Ioannis Ntzoufras, Professor

Department of Statistics, AUEB



2. Describing Data



- Organizing data to data frames
- Types of variables
- Questionnaire construction and validation
- From collection to data analysis
- Descriptive measures for each type of variable
- Contingency tables

2. Describing Data

Organizing data to data frames



In order to import data to a statistical package we need to fully specify

- Study Unit (observation unit, subject, object)
(and its size = sample size)

[denoted by n]

- Variables (i.e. characteristics of each observation unit)

[their number is denoted by p]

2. Describing Data

Organizing data to data frames



When we know these two basic ingredients
we can import our data in matrix form with

- n rows and
- p columns
- Each row contains the data of one observation unit
- Each column contains the data of a variable

2. Describing Data

Organizing data to data frames

A simple example

Four receipts were randomly sampled from one book store. In every receipt the total value and the number of books sold was recorded:

	<u>Receipt</u>			
<i>Variable</i>	1	2	3	4
1. <i>Value (€)</i>	42	52	48	58
2. <i>Book number</i>	4	5	4	3

2. Describing Data

Organizing data to data frames

A simple example

- *Observation unit: RECEIPT*
- *Sample size : $n=4$ receipts*
- *$p=2$ variables – characteristics:*
 - *Value in Euros*
 - *Number of books*
- *Data matrix* $\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$

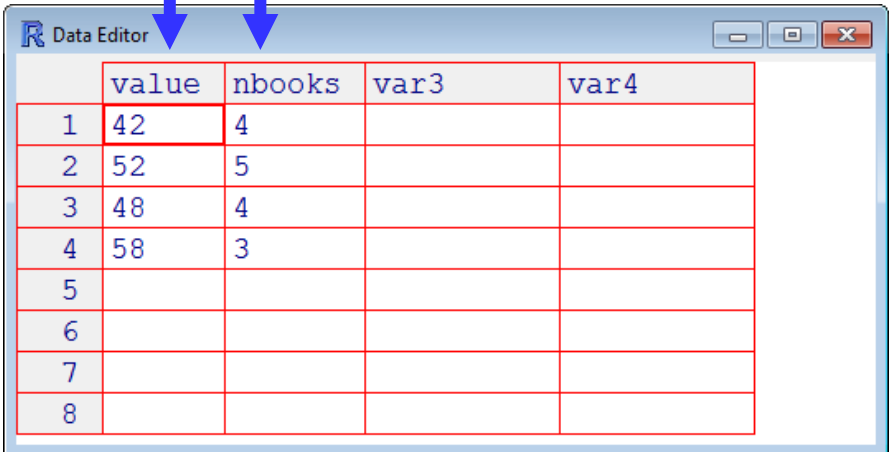
2. Describing Data

Organizing data to data frames

A simple example: Specification in R

```
ex1 <- data.frame(  
  value=c(42,52,48,58),  
  nbooks=c(4,5,4,3) )  
ex1<-edit(ex1)
```

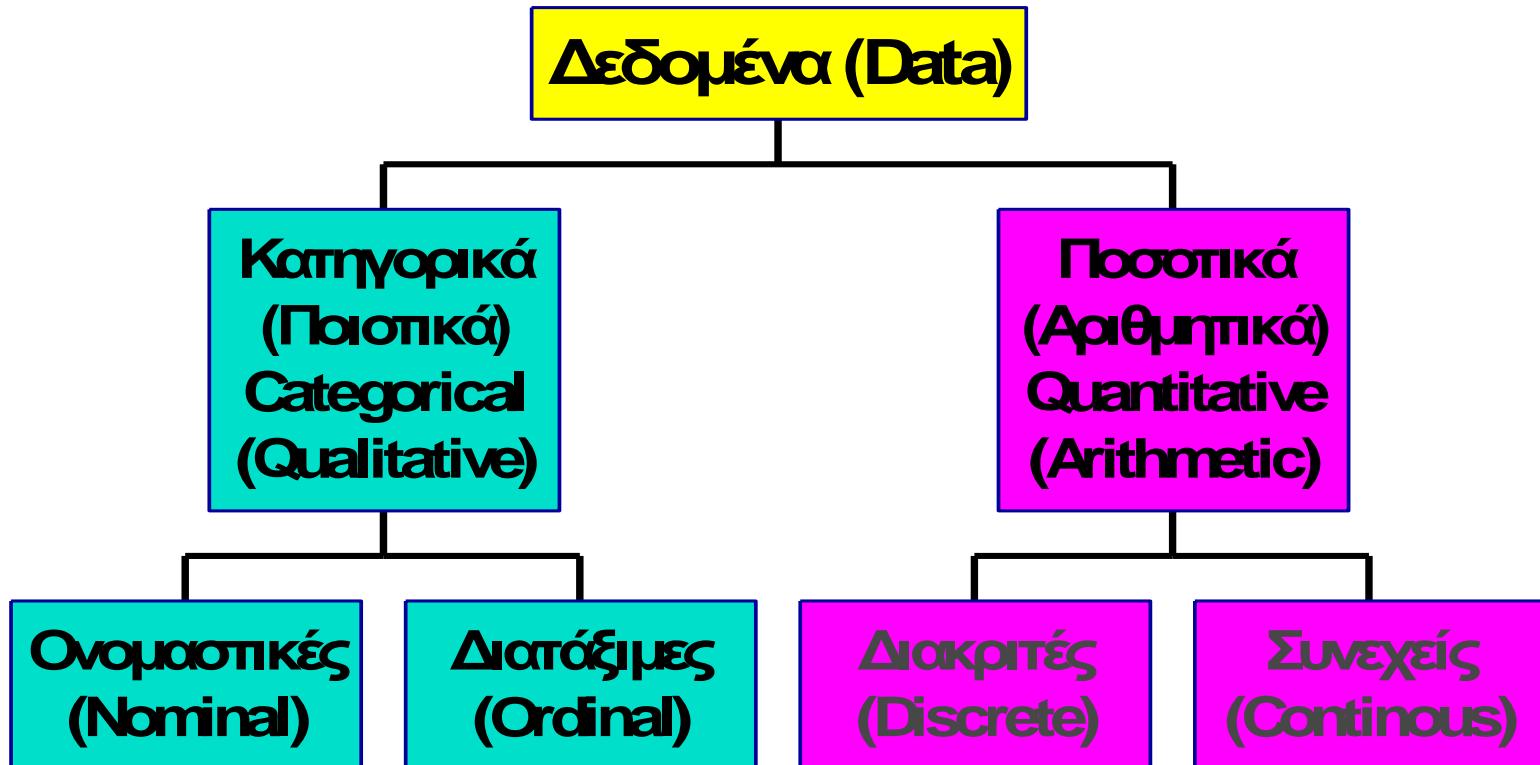
$$\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$



	value	nbooks	var3	var4
1	42	4		
2	52	5		
3	48	4		
4	58	3		
5				
6				
7				
8				

2. Describing Data

Type of variables



2. Describing Data

Quantitative variables

- Frequency tabulation
- Central location (mean, median, mode)
A typical/ordinary/average person
- Variance (standard deviation, IQR, R)
Homogeneity or divergence in groups
Close or open minded societies
- Relative location (Q_1 , Q_3 , $P_{0.25}$, $P_{0.975}$)
The best and the worst?
- Shape of distribution (kurtosis, skewness, symmetry)

2. Describing Data

Quantitative variables



Measures of Central location (mean, median, mode)

- What do we try to describe?
 - The center of the distribution?
 - A typical/ordinary/average/mediocre person/study unit
- What is the mean?
 - Is the mean always fair/descriptive enough of the average level?
- What is the median? [denoted by M]
- What is the mode?

2. Describing Data

Quantitative variables

Measures of Variability

- What do we try to describe?
 - Variability
 - Homogeneity
 - Risk
 - Uncertainty
- What is the variance?
 - Mean square distance from the mean
 - Measured in the squared units

2. Describing Data

Quantitative variables

Measures of Variability

- The standard deviation
 - Square root of variance
 - Measured in the same unit as the original variable
 - Use the normal distribution to get intervals and probabilities (mean \pm k SD)
 - Compare with mean (and obtain $CV=SD/\text{mean}$)
- The interquartile range: $IQR=Q_3-Q_1$
 - It is the range of the observations lying in 50% center of the distribution
- The median absolute deviation: MAD is the median of the absolute distance from the median

2. Describing Data

Quantitative variables

Measures of relative location

- Quantiles or percentiles
 - Indicates the value below which a given % of observations fall.
- Quantiles: Q_1 & M & Q_3
 - They split the data in 4 groups of (approx.) equal size
 - They are the 25%, 50% and 75% quantiles/percentiles
- Why quantiles/percentiles?
 - Society is also interested on the extremes
 - Which is the grade needed to enter the top 5% and get a scholarship?
 - What about sports, science or earthquakes?

2. Describing Data

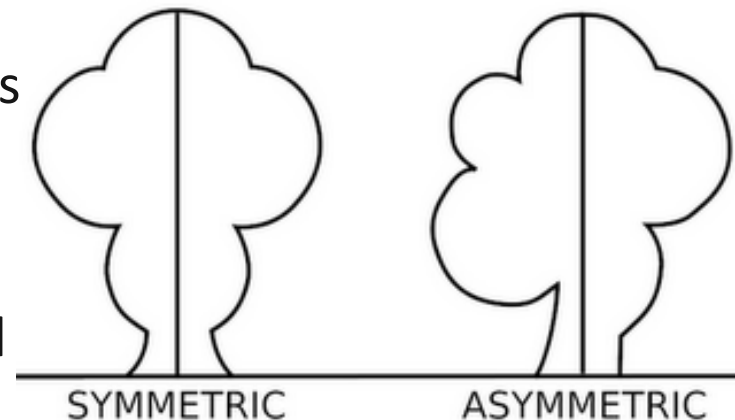
Quantitative variables

Measures of shape – symmetry or skewness & kurtosis

- Measures of shape are of interest in order to know
 - If the distribution is normal or close to normality
 - The behavior of the mean
 - The behavior of the extremes/outliers

Symmetry/skewness

- Measures whether values below and above the mean (probabilistically) behave (appear) in the same way

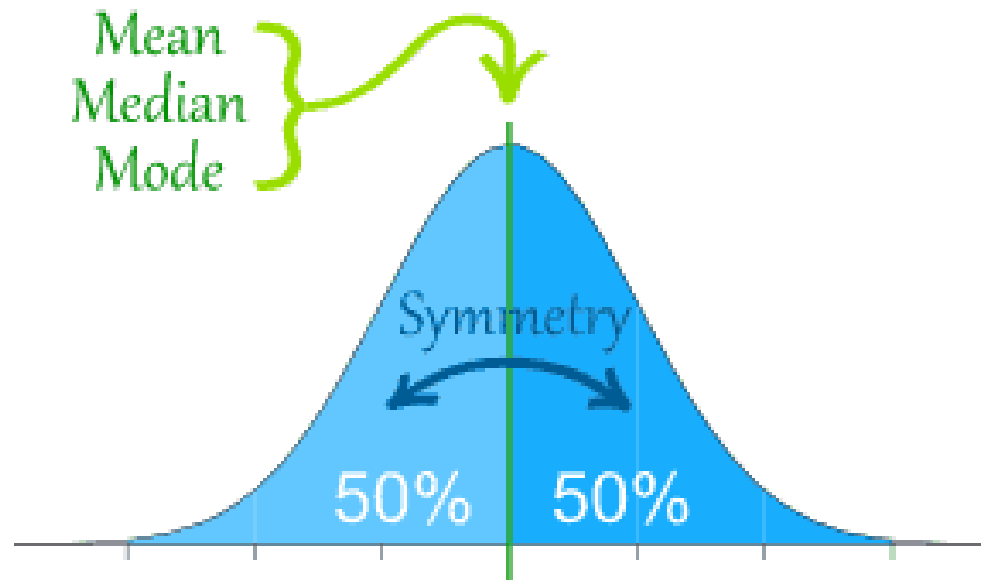


2. Describing Data

Quantitative variables

Symmetry/skewness

- Measures whether values below and above the mean (probabilistically) behave (appear) in the same way
- Mean – Mode and Median are the same



2. Describing Data

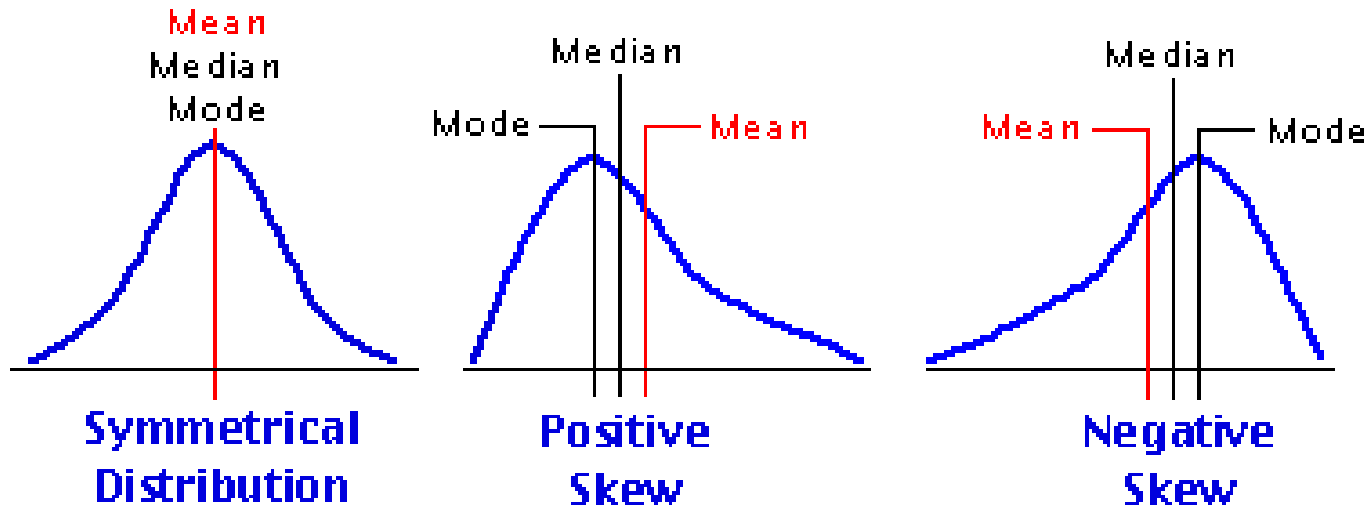
Quantitative variables



$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$

Symmetry/skewness

- Measures whether values below and above the mean (probabilistically) behave (appear) in the same way
- Mean – Mode and Median are the same



2. Describing Data

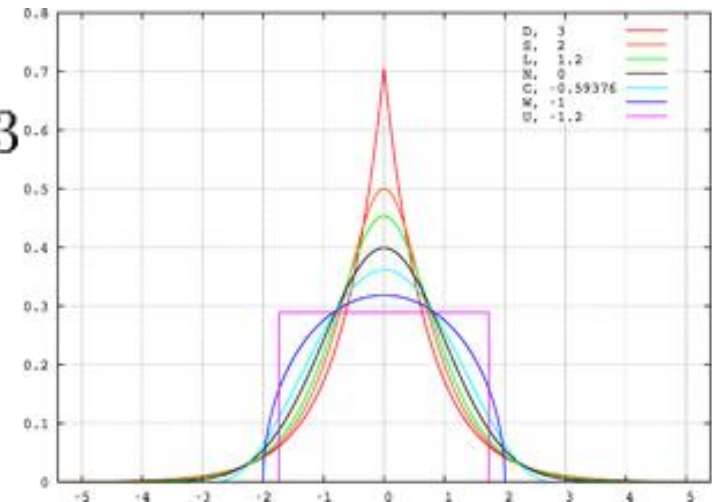
Quantitative variables



Kurtosis

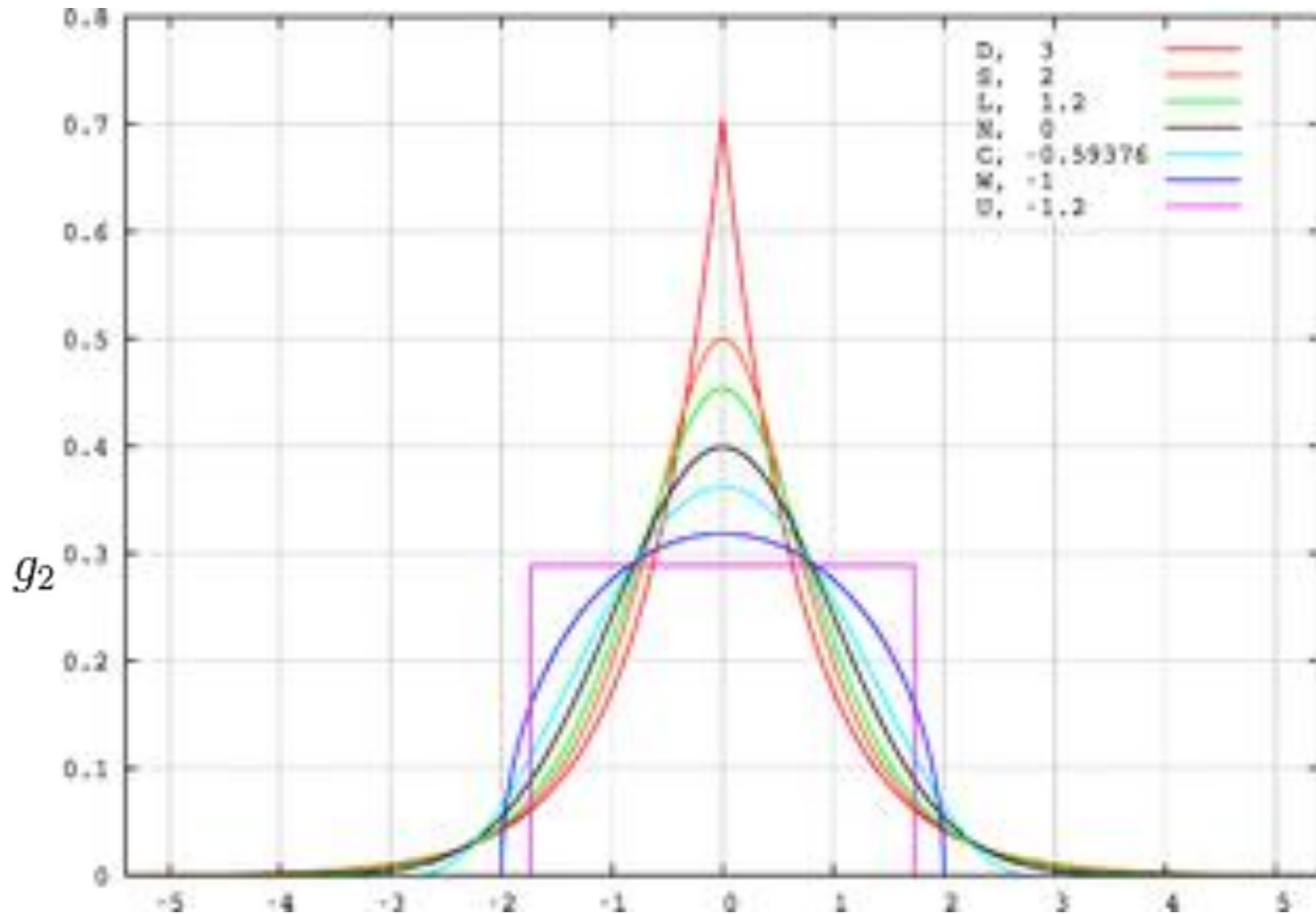
- is any measure of the "peakedness" (width of peak) of the distribution of real-valued random variables.
- It also measures tail weight, and lack of shoulders (distribution primarily peak and tails, not in between)

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3$$



2. Describing Data

Quantitative variables



2. Describing Data

Quantitative variables



Example: Baby boom dataset

SOURCE: The data appeared in the Brisbane newspaper `_The Sunday Mail_` on December 21, 1997.

Variables: Time, Sex (1 = girl, 2 = boy) , weight of and number of minutes after midnight for 44 baby births.

The dataset contains data for 44 babies born in one 24-hour period at a Brisbane, Australia, hospital. Also included is the number of minutes since midnight for each birth.

2. Describing Data

Example – reading the data in R



File: babyboom.dat

```
| 0005      1      3837      5
  0104      1      3334      64
  0118      2      3554      78
  0155      2      3838     115
  0257      2      3625     177
  0405      1      2208     245
  0407      1      1745     247
  0422      2      2846     262
```

```
> babyboom <-
read.table('babyboom.dat')
> babyboom[1:10,]
  V1 V2  V3  V4
1   5  1 3837  5
2 104  1 3334 64
3 118  2 3554 78
4 155  2 3838 115
5 257  2 3625 177
6 405  1 2208 245
7 407  1 1745 247
8 422  2 2846 262
9 431  2 3166 271
10 708  2 3520 428
> names(babyboom)
[1] "V1" "V2" "V3" "V4"
```

2. Describing Data

Example – reading the data in R

```
> names(babyboom)<-c( 'timebirth', 'gender', 'weight', 'min.after.mid' )
> babyboom$gender
[1] 1 1 2 2 2 1 1 2 2 2 2 2 1 1 2 1 1 2 2 2 2 1 1 1 1 2 2 2 1 2
[31] 1 2 2 2 2 2 1 2 2 2 2 1 1 1
> babyboom$g
[1] 1 1 2 2 2 1 1 2 2 2 2 2 1 1 2 1 1 2 2 2 2 1 1 1 1 2 2 2 1 2
[31] 1 2 2 2 2 2 1 2 2 2 2 1 1 1
> is.factor(babyboom$g)
[1] FALSE
> babyboom$gender <- factor( babyboom$gender, labels=c('girl', 'boy'))
> babyboom$g
[1] girl girl boy boy boy girl girl boy boy boy boy boy
[13] girl girl boy girl girl boy boy boy boy girl girl girl
[25] girl boy boy boy girl boy girl boy boy boy boy boy
[37] girl boy boy boy boy girl girl girl
Levels: girl boy
> mode(babyboom)
[1] "list"
> class(babyboom)
[1] "data.frame"
```

2. Describing Data

Example – Frequency tabulations



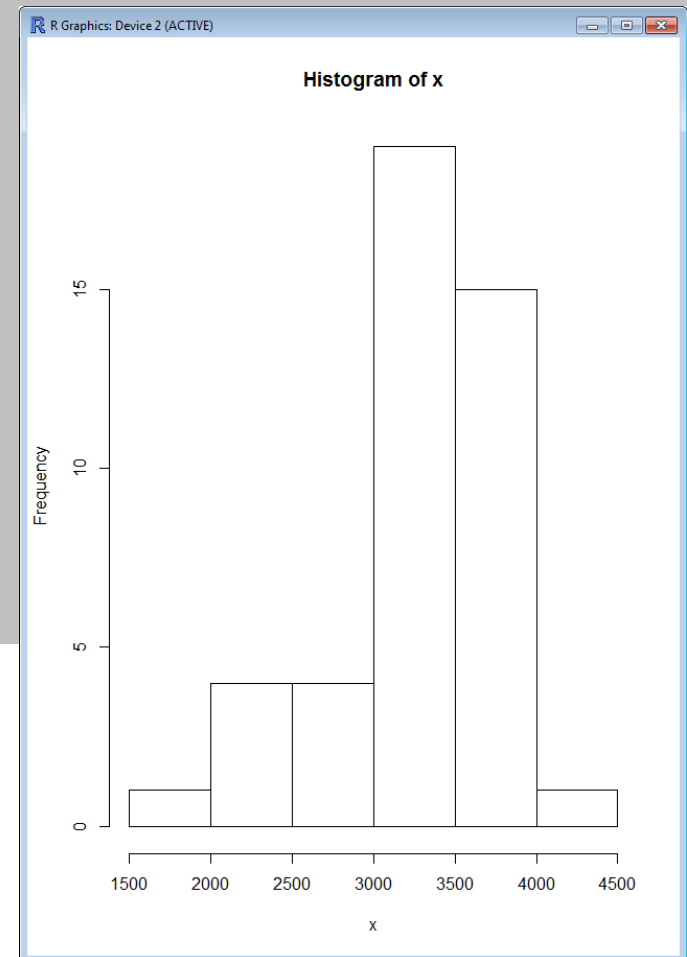
```
> x<-babyboom$weight
> factorx <- factor(cut(x, breaks=nclass.Sturges(x)))
> table(factorx)
factorx
(1.74e+03,2.09e+03] (2.09e+03,2.44e+03] (2.44e+03,2.78e+03]
(2.78e+03,3.13e+03] (3.13e+03,3.47e+03]
      1          4          2          4          15
(3.47e+03,3.82e+03] (3.82e+03,4.16e+03]
      13          5
> as.matrix(table(factorx))
      [,1]
(1.74e+03,2.09e+03]  1
(2.09e+03,2.44e+03]  4
(2.44e+03,2.78e+03]  2
(2.78e+03,3.13e+03]  4
(3.13e+03,3.47e+03] 15
(3.47e+03,3.82e+03] 13
(3.82e+03,4.16e+03]  5
```

2. Describing Data

Example – Frequency tabulations



```
> factorx <- factor(cut(x, breaks=nclass.Sturges(x), dig.lab=5))  
> as.matrix(table(factorx))  
      [,1]  
(1742.6,2090.3] 1  
(2090.3,2435.6] 4  
(2435.6,2780.9] 2  
(2780.9,3126.1] 4  
(3126.1,3471.4] 15  
(3471.4,3816.7] 13  
(3816.7,4164.4] 5  
> hist(x)
```



2. Describing Data

Example – Frequency tabulations

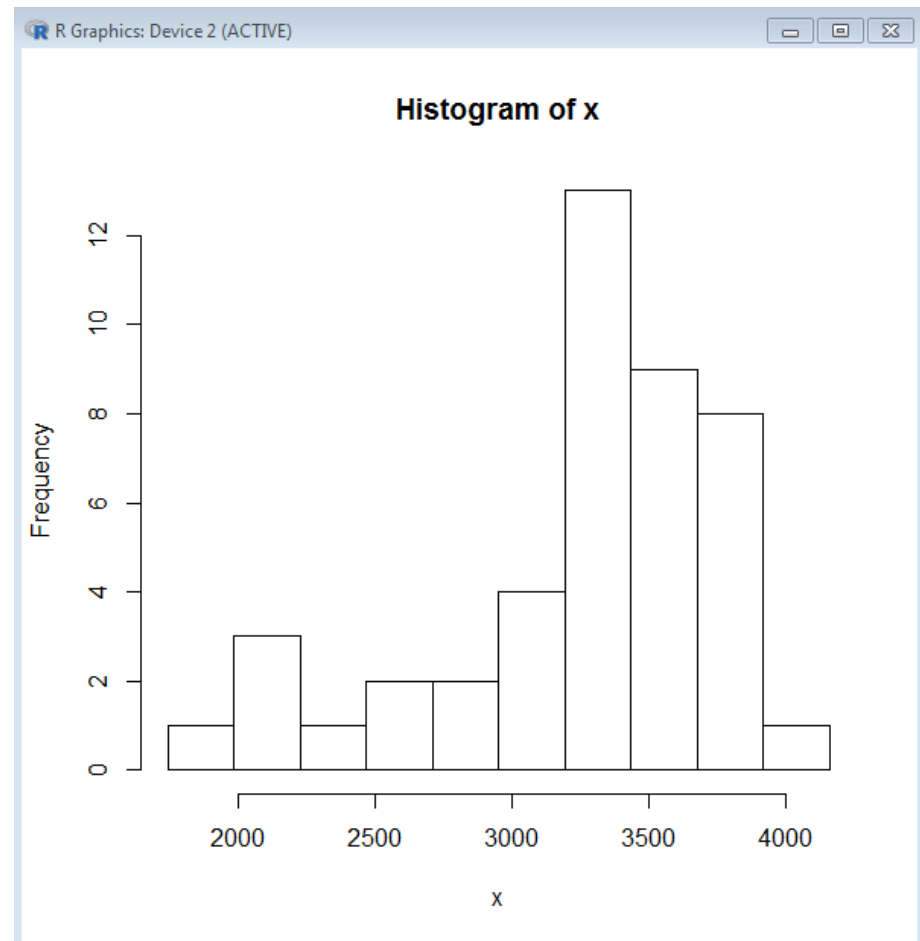
```
factorx <- factor(cut(x, breaks=nclass.scott(x), dig.lab=5))  
factorx <- factor(cut(x, breaks=nclass.FD(x), dig.lab=5))  
  
nclasses <- 10  
step <- (max(x)-min(x))/nclasses  
factorx <- factor(cut(x, breaks=seq( min(x), max(x), step ),  
include.lowest = TRUE, dig.lab=5))  
as.matrix(table(factorx))  
hist(x, breaks=seq( min(x), max(x), step ) )
```


2. Describing Data

Example – Frequency tabulations



```
> nclasses <- 10
> step <- (max(x)-min(x))/nclasses
> factorx <- factor(cut(x, breaks=seq( min(x), max(x), step), include.lowest = TRUE, dig.lab=5))
> as.matrix(table(factorx))
      [,1]
[1745,1986.7] 1
[1986.7,2228.4] 3
[2228.4,2470.1] 1
[2470.1,2711.8] 2
[2711.8,2953.5] 2
[2953.5,3195.2] 4
[3195.2,3436.9] 13
[3436.9,3678.6] 9
[3678.6,3920.3] 8
[3920.3,4162] 1
> hist(x, breaks=seq( min(x), max(x), step ) )
```



2. Describing Data

Example – Frequency tabulations



```
> nclasses <- 10
> step <- (max(x)-min(x))/nclasses
> factorx <- factor(cut(x, breaks=seq( min(x), max(x), step), include.lowest
= TRUE, dig.lab=5))
> #Tabulate and turn into data.frame
> Freq <- table(factorx)
> rel.Freq <- prop.table(Freq)
> xout <- data.frame(Freq=as.numeric(Freq), cum.Freq = cumsum(Freq),
rel.Freq = as.numeric(rel.Freq), cum.rel.Freq=cumsum(rel.Freq))
> round(xout,3)
```

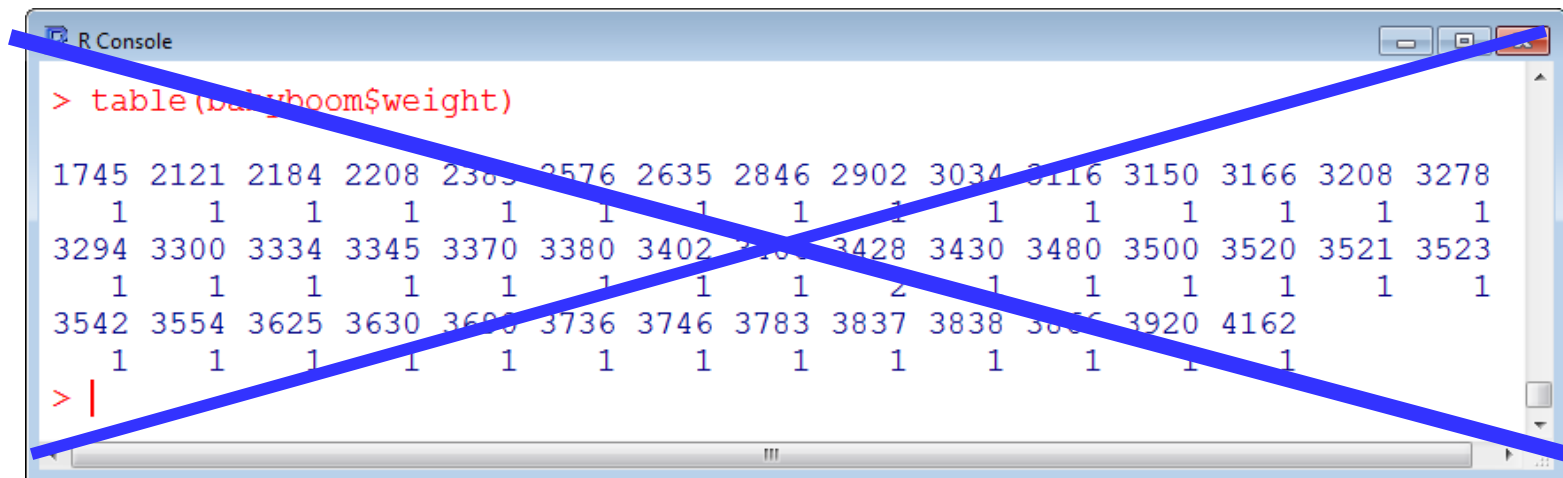
	Freq	cum.Freq	rel.Freq	cum.rel.Freq
[1745,1986.7]	1	1	0.023	0.023
(1986.7,2228.4]	3	4	0.068	0.091
(2228.4,2470.1]	1	5	0.023	0.114
(2470.1,2711.8]	2	7	0.045	0.159
(2711.8,2953.5]	2	9	0.045	0.205
(2953.5,3195.2]	4	13	0.091	0.295
(3195.2,3436.9]	13	26	0.295	0.591
(3436.9,3678.6]	9	35	0.205	0.795
(3678.6,3920.3]	8	43	0.182	0.977
(3920.3,4162]	1	44	0.023	1.000

2. Describing Data

Example – Frequency tabulations

BE CAREFUL

- Do not apply frequency tabulation directly on variables (vectors in R)
- Only on tables!
- For continuous variables you need to use the cut command to change it to factors



```
> table(balghoom$weight)
1745 2121 2184 2208 2385 2576 2635 2846 2902 3034 3116 3150 3166 3208 3278
 1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
3294 3300 3334 3345 3370 3380 3402 3410 3428 3430 3480 3500 3520 3521 3523
 1    1    1    1    1    1    1    1    2    1    1    1    1    1    1
3542 3554 3625 3630 3690 3736 3746 3783 3837 3838 3850 3920 4162
 1    1    1    1    1    1    1    1    1    1    1    1    1    1
> |
```

2. Describing Data

Example – Frequency tabulations

sjPlot/ sjmisc libraries

Create excellent output in html format (editable by word and open office)

```
library(sjPlot)
library(sjmisc)
frq(factorx, title="Birth
Weight", out = "v")
```

Birth Weight

<i>val</i>	<i>label</i>	<i>frq</i>	<i>raw.prc</i>	<i>valid.prc</i>	<i>cum.prc</i>
[1745,1986.7]		1	2.27	2.27	2.27
(1986.7,2228.4]		3	6.82	6.82	9.09
(2228.4,2470.1]		1	2.27	2.27	11.36
(2470.1,2711.8]		2	4.55	4.55	15.91
(2711.8,2953.5]		2	4.55	4.55	20.45
(2953.5,3195.2]		4	9.09	9.09	29.55
(3195.2,3436.9]		13	29.55	29.55	59.09
(3436.9,3678.6]		9	20.45	20.45	79.55
(3678.6,3920.3]		8	18.18	18.18	97.73
(3920.3,4162]		1	2.27	2.27	100
NA	NA	0	0	NA	NA

total N=44 · valid N=44 · \bar{x} =6.75 · σ =2.18

2. Describing Data

Example – Frequency tabulations

sjPlot/sjmisc libraries

Create excellent output in html format (editable by word and open office)

```
library(sjPlot)
library(sjmisc)
frq(babyboom$weight,
title="Birth Weight", out =
"v")
```

Birth Weight

<i>val</i>	<i>label</i>	<i>frq</i>	<i>raw.prc</i>	<i>valid.prc</i>	<i>cum.prc</i>
1	1740-1989	1	2.27	2.27	2.27
2	1990-2229	3	6.82	6.82	9.09
3	2230-2469	1	2.27	2.27	11.36
4	2470-2709	2	4.55	4.55	15.91
5	2710-2959	2	4.55	4.55	20.45
6	2960-3199	4	9.09	9.09	29.55
7	3200-3439	13	29.55	29.55	59.09
8	3440-3679	9	20.45	20.45	79.55
9	3680-3919	8	18.18	18.18	97.73
10	3920-4159	1	2.27	2.27	100
NA	NA	0	0	NA	NA

total N=44 · valid N=44 · \bar{x} =3275.95 · σ =528.03

2. Describing Data

Example – Descriptive measures



Descriptives in R

- mean, median
- var, sd, mad, IQR
- `quantile(x, probs=c(0.25, 0.5, 0.75))`
- range, min, max
- `skew(x)`, `kurtosis(x)` in package 'psych'

All summaries together

- `summary(dataframe)`
- `library(psych) => describe & describe.by`

2. Describing Data

Example – Descriptive measures



Descriptives in R – using vectors

```
x<-babyboom$weight  
mean(x)  
median(x)  
  
var(x)  
sd(x)  
mad(x)  
IQR(x)  
  
range(x)  
min(x)  
max(x)  
pp=c( 0.005, 0.025,  
seq(0.05,0.95,0.05), 0.975, 0.995 )
```

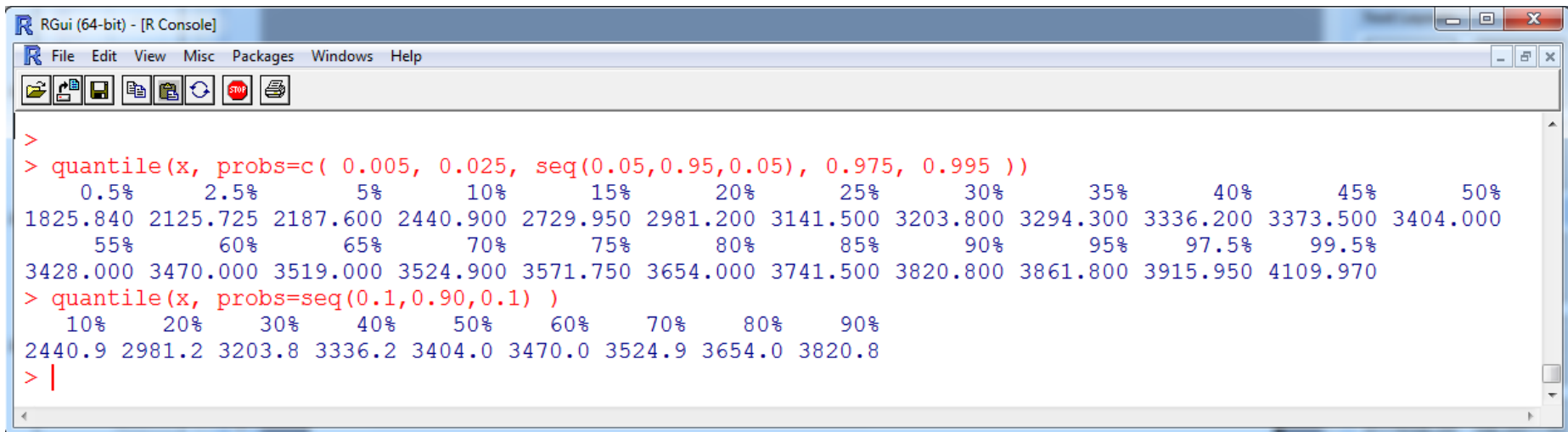
```
RGui (64-bit) - [R Co...  
File Edit View Misc Packages  
Windows Help  
< > < > < > < > < > < > < >  
> mean(x)  
[1] 3275.955  
> median(x)  
[1] 3404  
> var(x)  
[1] 278818.3  
> sd(x)  
[1] 528.0325  
> mad(x)  
[1] 343.9632  
> IQR(x)  
[1] 430.25  
>  
> range(x)  
[1] 1745 4162  
> min(x)  
[1] 1745  
> max(x)  
[1] 4162  
>
```

2. Describing Data

Example – Descriptive measures

Quantiles in R using vectors

```
quantile(x, probs=pp)  
quantile(x, probs=seq(0.1,0.90,0.1) )
```



```
RGui (64-bit) - [R Console]  
File Edit View Misc Packages Windows Help  
>  
> quantile(x, probs=c( 0.005, 0.025, seq(0.05,0.95,0.05), 0.975, 0.995 ))  
 0.5%   2.5%   5%   10%   15%   20%   25%   30%   35%   40%   45%   50%  
1825.840 2125.725 2187.600 2440.900 2729.950 2981.200 3141.500 3203.800 3294.300 3336.200 3373.500 3404.000  
 55%   60%   65%   70%   75%   80%   85%   90%   95%   97.5%  99.5%  
3428.000 3470.000 3519.000 3524.900 3571.750 3654.000 3741.500 3820.800 3861.800 3915.950 4109.970  
> quantile(x, probs=seq(0.1,0.90,0.1) )  
 10%   20%   30%   40%   50%   60%   70%   80%   90%  
2440.9 2981.2 3203.8 3336.2 3404.0 3470.0 3524.9 3654.0 3820.8  
> |
```


2. Describing Data

Example – Descriptive measures

Descriptives in R – using data frames

```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

>
> mean(x)
[1] NA
Warning message:
In mean.default(x) : argument is not numeric or logical: returning NA
> median(x)
Error in median.default(x) : need numeric data
> var(x)
      timebirth gender  weight min.after.mid
timebirth  477072.79   NA  27585.67   287333.26
gender           NA   NA         NA           NA
weight      27585.67   NA 278818.28   17491.38
min.after.mid 287333.26   NA  17491.38   173111.69
Warning message:
In var(x) : NAs introduced by coercion
> sd(x)
Error in is.data.frame(x) :
 (list) object cannot be coerced to type 'double'
> mad(x)
Error in median.default(x) : need numeric data
> IQR(x)
Error in quantile(as.numeric(x), c(0.25, 0.75), na.rm = na.rm, names = FALSE, :
 (list) object cannot be coerced to type 'double'
>
> range(x)
Error in FUN(X[[1L]], ...) :
 only defined on a data frame with all numeric variables
> min(x)
Error in FUN(X[[1L]], ...) :
 only defined on a data frame with all numeric variables
> max(x)
Error in FUN(X[[1L]], ...) :
 only defined on a data frame with all numeric variables
>
> quantile(x, probs=c( 0.005, 0.025, seq(0.05,0.95,0.05), 0.975, 0.995 ))
Error in `[.data.frame`(x, order(x, na.last = na.last, decreasing = decreasing)) :
 undefined columns selected
> quantile(x, probs=seq(0.1,0.90,0.1) )
Error in `[.data.frame`(x, order(x, na.last = na.last, decreasing = decreasing)) :
 undefined columns selected
> |
```

2. Describing Data

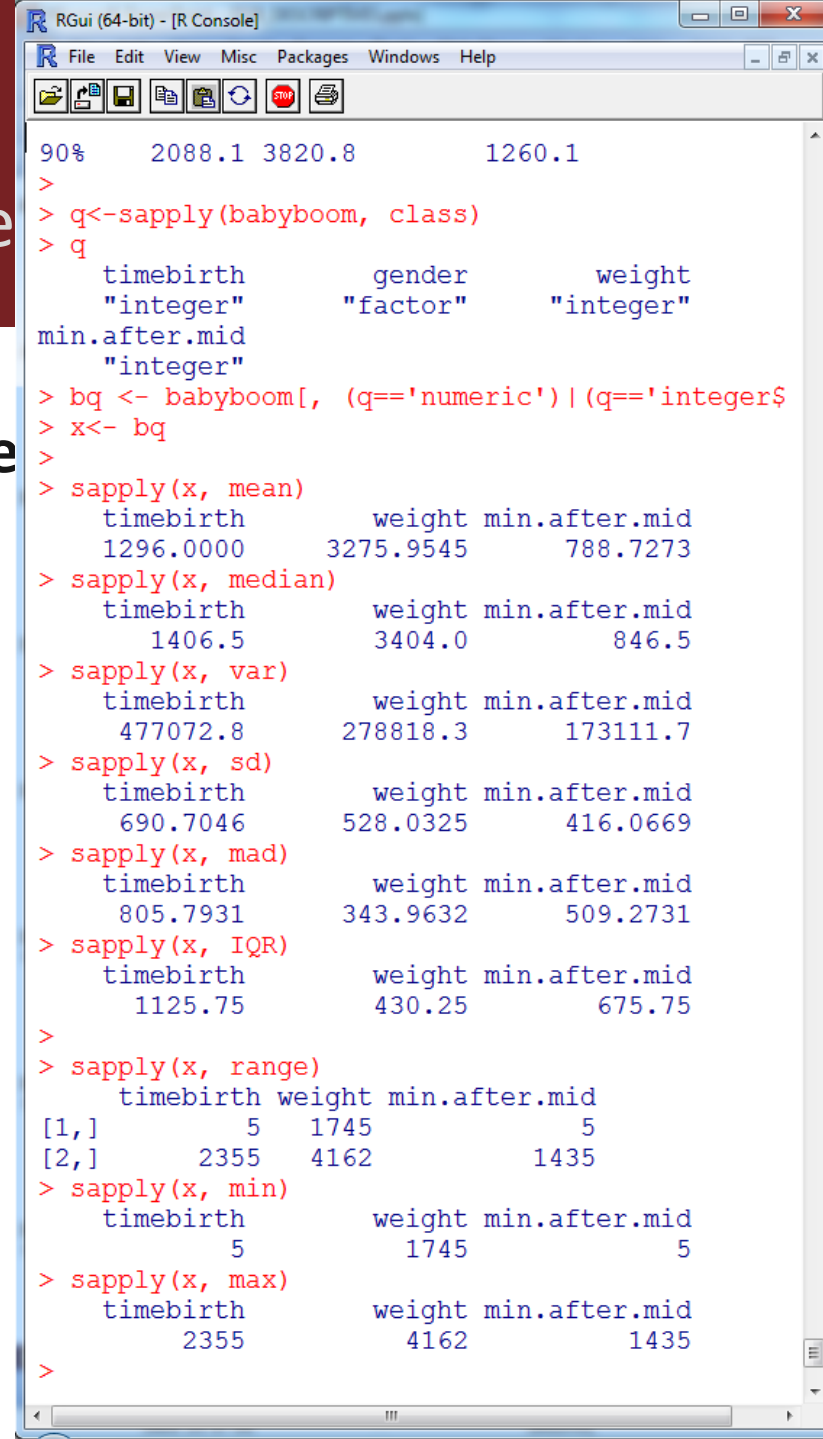
Example – Descriptive measure

Descriptives in R – using data frame

```
q<-sapply(babyboom, class)
q
bq <- babyboom[,
(q=='numeric')|(q=='integer')]
x<- bq
```

```
sapply(x, mean)
sapply(x, median)
sapply(x, var)
sapply(x, sd)
sapply(x, mad)
sapply(x, IQR)
```

```
sapply(x, range)
sapply(x, min)
sapply(x, max)
```



```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
[Icons]
90% 2088.1 3820.8 1260.1
>
> q<-sapply(babyboom, class)
> q
  timebirth      gender      weight
  "integer"    "factor"    "integer"
min.after.mid
  "integer"
> bq <- babyboom[, (q=='numeric')|(q=='integer')]
> x<- bq
>
> sapply(x, mean)
  timebirth      weight min.after.mid
1296.0000    3275.9545     788.7273
> sapply(x, median)
  timebirth      weight min.after.mid
 1406.5        3404.0      846.5
> sapply(x, var)
  timebirth      weight min.after.mid
477072.8        278818.3    173111.7
> sapply(x, sd)
  timebirth      weight min.after.mid
 690.7046        528.0325    416.0669
> sapply(x, mad)
  timebirth      weight min.after.mid
 805.7931        343.9632    509.2731
> sapply(x, IQR)
  timebirth      weight min.after.mid
 1125.75         430.25     675.75
>
> sapply(x, range)
  timebirth weight min.after.mid
[1,]      5    1745      5
[2,]    2355    4162    1435
> sapply(x, min)
  timebirth      weight min.after.mid
      5          1745      5
> sapply(x, max)
  timebirth      weight min.after.mid
    2355         4162    1435
>
```

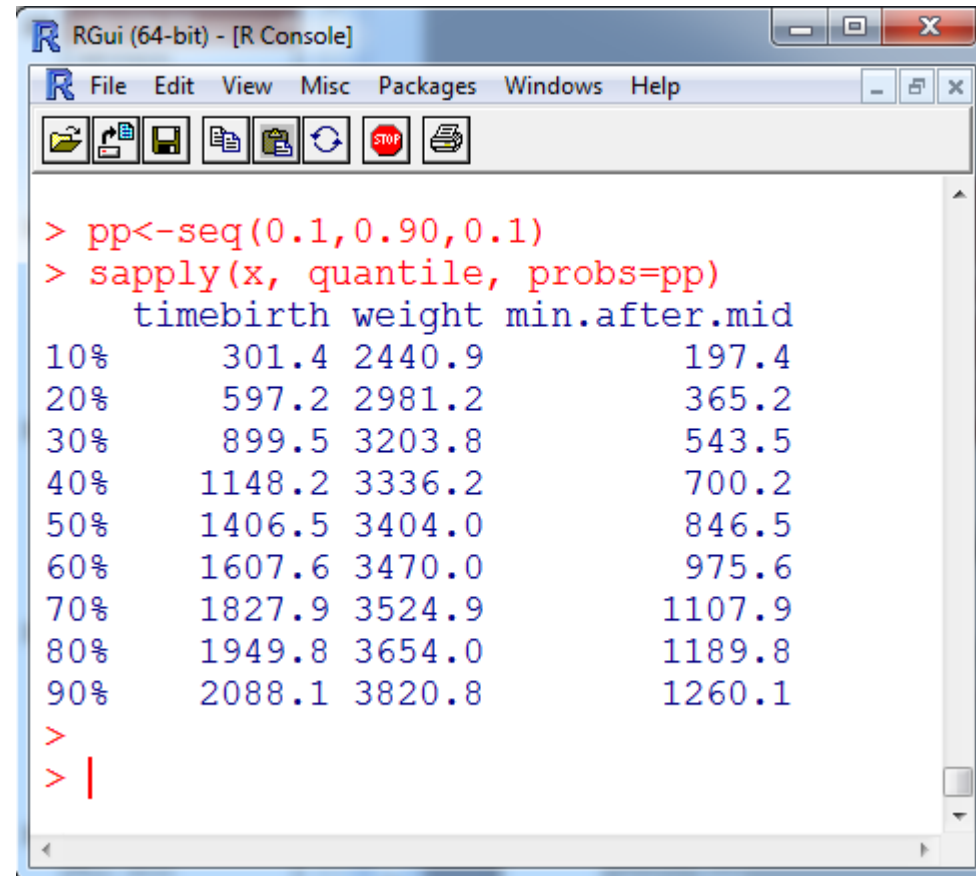
2. Describing Data

Example – Descriptive measures

Quantiles in R using dataframes

```
q<-sapply(babyboom, class)
q
bq <- babyboom[,
(q=='numeric')|(q=='integer')]
x<- bq
```

```
pp<-seq(0.1,0.90,0.1)
sapply(x, quantile, probs=pp)
```



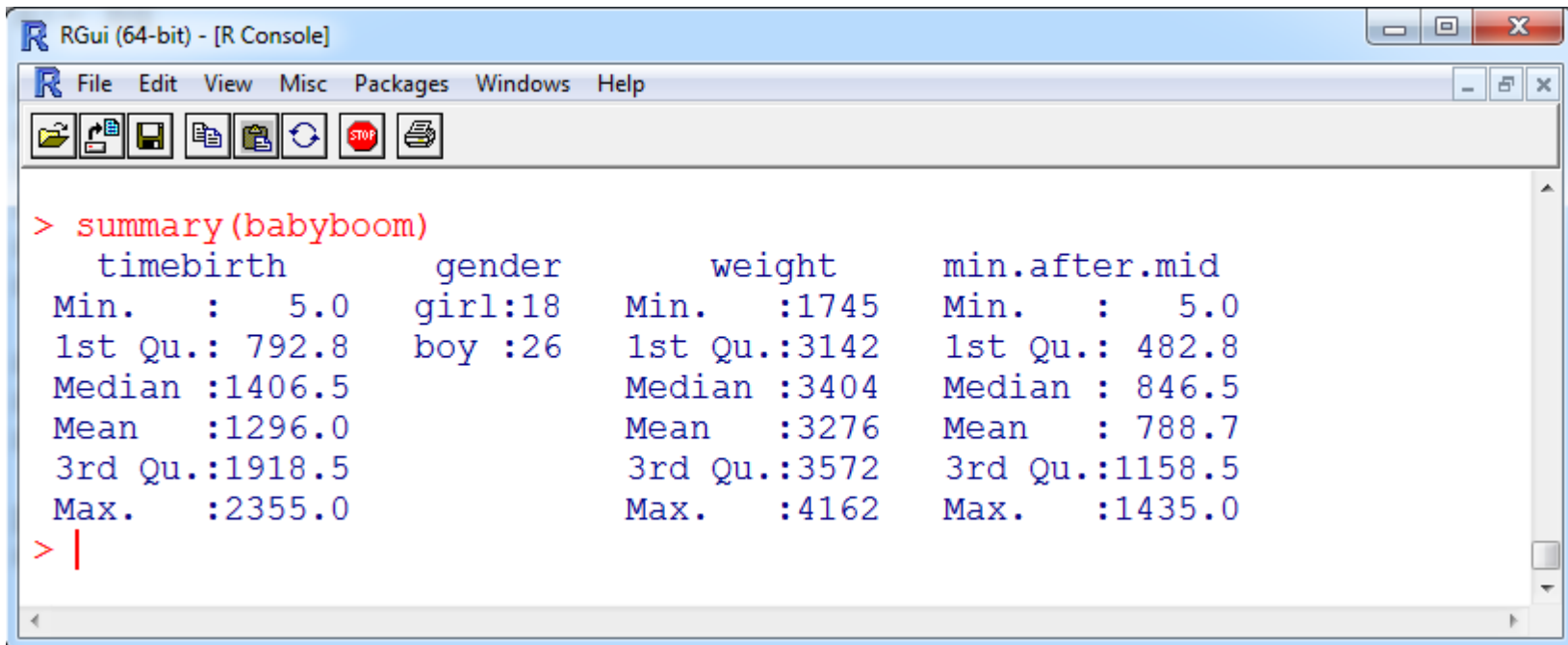
```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
[Icons]
> pp<-seq(0.1,0.90,0.1)
> sapply(x, quantile, probs=pp)
      timebirth weight min.after.mid
10%      301.4  2440.9      197.4
20%      597.2  2981.2      365.2
30%      899.5  3203.8      543.5
40%     1148.2  3336.2      700.2
50%     1406.5  3404.0      846.5
60%     1607.6  3470.0      975.6
70%     1827.9  3524.9     1107.9
80%     1949.8  3654.0     1189.8
90%     2088.1  3820.8     1260.1
>
> |
```

2. Describing Data

Example – Descriptive measures

Summary statistics of dataframes

```
summary(babyboom)
```



```
> summary(babyboom)
  timebirth      gender      weight  min.after.mid
Min.   :    5.0  girl:18  Min.   :1745  Min.   :    5.0
1st Qu.:  792.8  boy :26   1st Qu.:3142  1st Qu.:  482.8
Median : 1406.5                Median :3404  Median :  846.5
Mean   : 1296.0                Mean   :3276  Mean   :  788.7
3rd Qu.:1918.5                3rd Qu.:3572  3rd Qu.:1158.5
Max.   :2355.0                Max.   :4162  Max.   :1435.0
> |
```

2. Describing Data

Example – Descriptive measures



Summary statistics of dataframes

```
library(psych)
describe(babyboom)
```

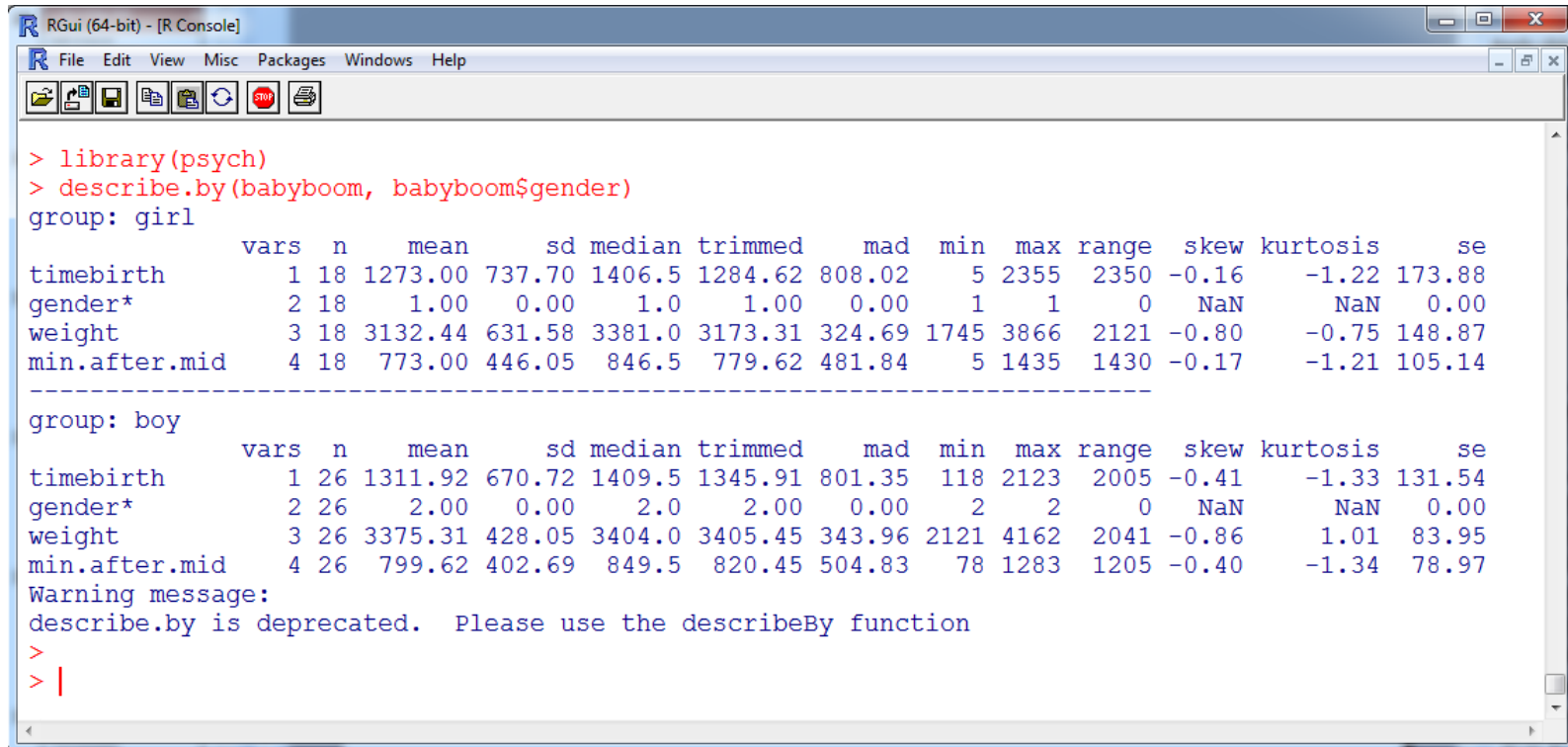
```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
> library(psych)
> describe(babyboom)
      vars  n   mean    sd median trimmed   mad  min  max range  skew kurtosis   se
timebirth  1 44 1296.00 690.70 1406.5 1322.78 805.79    5 2355 2350 -0.31   -1.18 104.13
gender*    2 44   1.59   0.50   2.0   1.61   0.00    1  2    1 -0.36   -1.91  0.07
weight     3 44 3275.95 528.03 3404.0 3336.06 343.96 1745 4162 2417 -1.08    0.64  79.60
min.after.mid 4 44  788.73 416.07  846.5  805.00 509.27    5 1435 1430 -0.31   -1.18  62.72
> |
```

2. Describing Data

Example – Descriptive measures

Summary statistics of dataframes

```
library(psych)
describe.by(babyboom, babyboom$gender)
```



```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
> library(psych)
> describe.by(babyboom, babyboom$gender)
group: girl
      vars  n   mean    sd median trimmed   mad  min  max range  skew kurtosis   se
timebirth  1 18 1273.00 737.70 1406.5 1284.62 808.02   5 2355 2350 -0.16   -1.22 173.88
gender*    2 18   1.00   0.00   1.0   1.00   0.00   1   1   0   NaN     NaN   0.00
weight     3 18 3132.44 631.58 3381.0 3173.31 324.69 1745 3866 2121 -0.80   -0.75 148.87
min.after.mid 4 18  773.00 446.05  846.5  779.62 481.84   5 1435 1430 -0.17   -1.21 105.14
-----
group: boy
      vars  n   mean    sd median trimmed   mad  min  max range  skew kurtosis   se
timebirth  1 26 1311.92 670.72 1409.5 1345.91 801.35  118 2123 2005 -0.41   -1.33 131.54
gender*    2 26   2.00   0.00   2.0   2.00   0.00   2   2   0   NaN     NaN   0.00
weight     3 26 3375.31 428.05 3404.0 3405.45 343.96 2121 4162 2041 -0.86    1.01  83.95
min.after.mid 4 26  799.62 402.69  849.5  820.45 504.83   78 1283 1205 -0.40   -1.34  78.97
Warning message:
describe.by is deprecated. Please use the describeBy function
>
> |
```

2. Describing Data

Categorical variables



Categorical variables – nominal or qualitative

- Frequency tabulation
- The mode

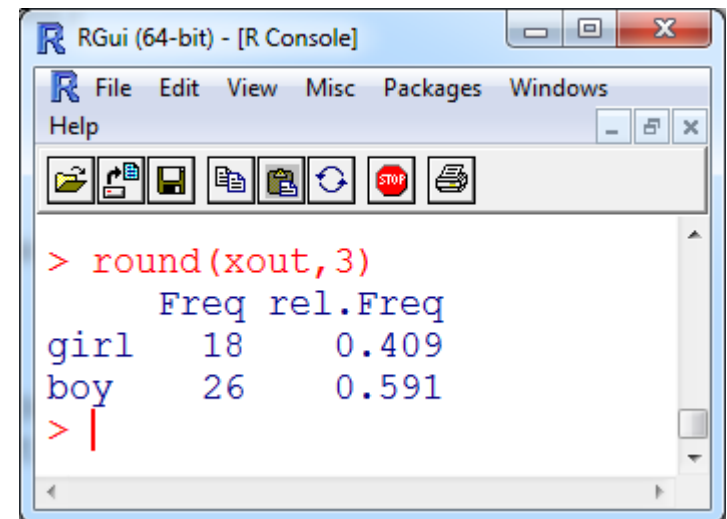
2. Describing Data

Categorical variables

Frequency tabulation in R

- Babyboom dataset
- Variable gender

```
x<-babyboom$gender
Freq <- table(x)
rel.Freq <- prop.table(Freq)
xout <- data.frame(Freq=as.numeric(Freq),
rel.Freq = as.numeric(rel.Freq))
row.names(xout) <- names(Freq)
round(xout,3)
```



```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows
Help
> round(xout,3)
      Freq rel.Freq
girl   18   0.409
boy   26   0.591
> |
```


2. Describing Data

Categorical variables

Frequency tabulation in R

- Babyboom dataset
- Variable gender

```
library(sjmisc)  
frq(babyboom$gender, title="Gender of newborn")
```

Οι Αθροιστικές
συχνότητες δεν
έχουν νόημα για
κατηγορικές

Gender		Gender of newborn					
<i>val</i>	<i>label</i>	<i>val</i>	<i>label</i>	<i>frq</i>	<i>raw.prc</i>	<i>valid.prc</i>	<i>cum.prc</i>
girl		girl		18	40.91	40.91	40.91
boy		boy		26	59.09	59.09	100
NA	NA	NA	NA	0	0	NA	NA
		<i>total N=44 · valid N=44 · $\bar{x}=1.59$ · $\sigma=0.50$</i>					

2. Describing Data

Categorical variables

Masticha Shop dataset

- Subsample from a customer satisfaction survey
- Sample size = 35

```
frq(masticha$residence, title="Τόπος Κατοικίας")
```

Τόπος Κατοικίας

<i>value</i>	<i>N</i>	<i>raw %</i>	<i>valid %</i>	<i>cumulative %</i>
Αττική	30	85.71	85.71	85.71
Θεσσαλονίκη	3	8.57	8.57	94.29
Άλλο-Επαρχία	2	5.71	5.71	100.00
missings	0	0.00		

total N=35 · valid N=35 · \bar{x} =1.20 · σ =0.53

2. Describing Data

Categorical variables

Masticha Shop dataset

- Subsample from a customer satisfaction survey
- Sample size = 35

```
frq(masticha$reason.of.visit, title="Λόγος Επίσκεψης")
```

Λόγος Επίσκεψης				
<i>value</i>	<i>N</i>	<i>raw %</i>	<i>valid %</i>	<i>cumulative %</i>
Για να αγοράσω σουβενίρ & δώρα	16	45.71	48.48	48.48
Από περιέργεια	7	20.00	21.21	69.70
Μου αρέσουν τα προϊόντα μαστίχας	9	25.71	27.27	96.97
Άλλο	1	2.86	3.03	100.00
missings	2	5.71		

total N=35 · valid N=33 · \bar{x} =1.85 · σ =0.94

2. Describing Data

Categorical variables

Masticha Shop dataset

- Finding the mode

```
x<-masticha$reason.of.visit  
tabx <- table(x)  
tabx[which(tabx == max(tabx))]
```

Λόγος Επίσκεψης

<i>value</i>	<i>N</i>	<i>raw %</i>	<i>valid %</i>	<i>cumulative %</i>
Για να αγοράσω σουβενίρ & δώρα	16	45.71	48.48	48.48
Από περιέργεια	7	20.00	21.21	69.70
Μου αρέσουν τα προϊόντα μαστίχας	9			
Άλλο	1			
missings	2			
<i>total</i>				

```
R Console  
> tabx <- table(x)  
> tabx  
x  
  Για να αγοράσω σουβενίρ & δώρα      Από περιέργεια  
  16                                     7  
  Μου αρέσουν τα προϊόντα μαστίχας      Άλλο  
  9                                       1  
> tabx[which(tabx == max(tabx))]  
Για να αγοράσω σουβενίρ & δώρα  
16  
> |
```

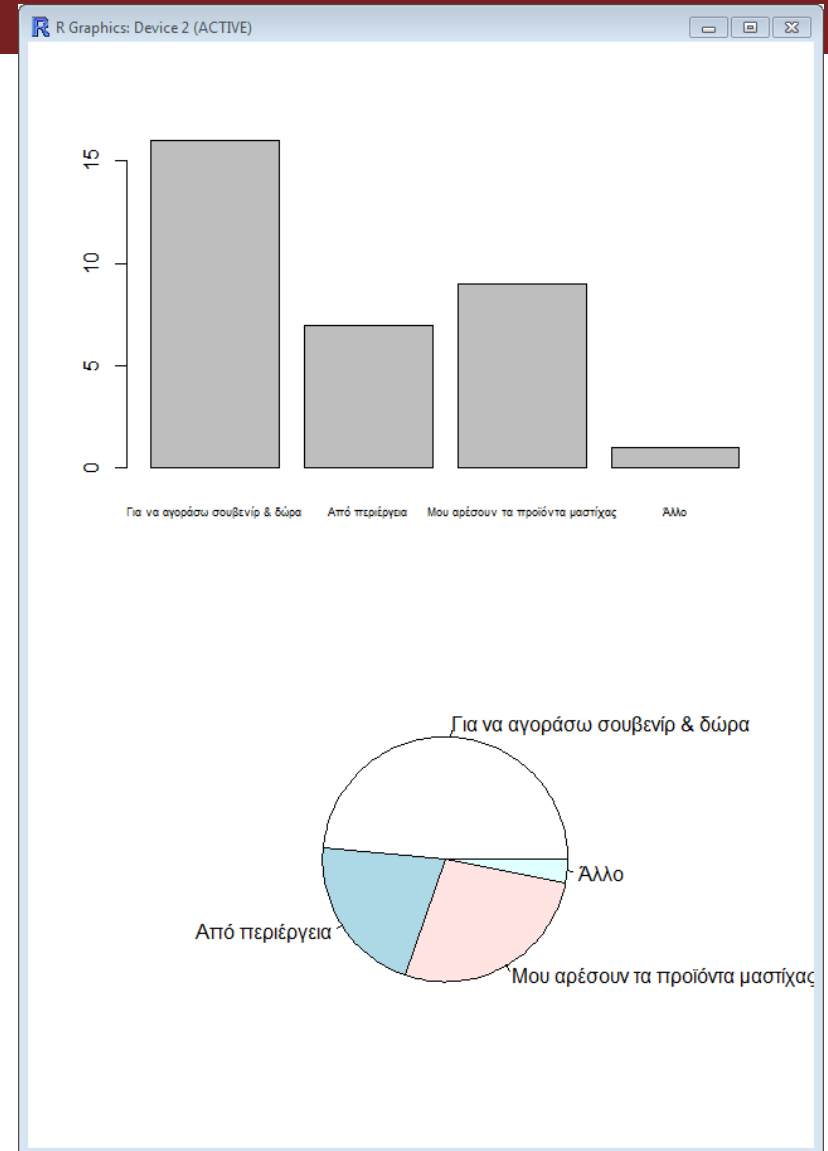
2. Describing Data

Categorical variables

Masticha Shop dataset

- Subsample from a customer satisfaction survey
- Sample size = 35

```
par(mfrow=c(2,1))  
barplot(table(masticha$reason.of.visit))  
pie(table(masticha$reason.of.visit))
```

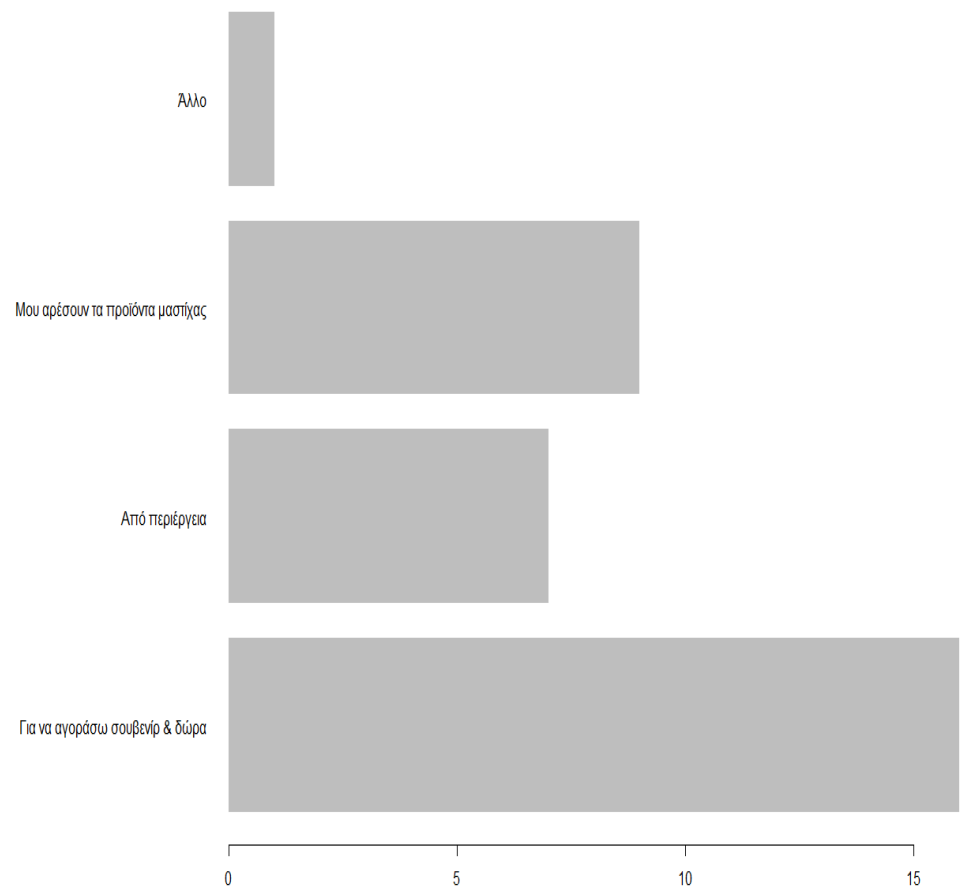


2. Describing Data

Categorical variables

Masticha Shop dataset

```
par(mar = c(5, 16, 4, 2))  
barplot(table(masticha$reason.of.vi  
sit), horiz=TRUE, las=1,  
border=NA)
```

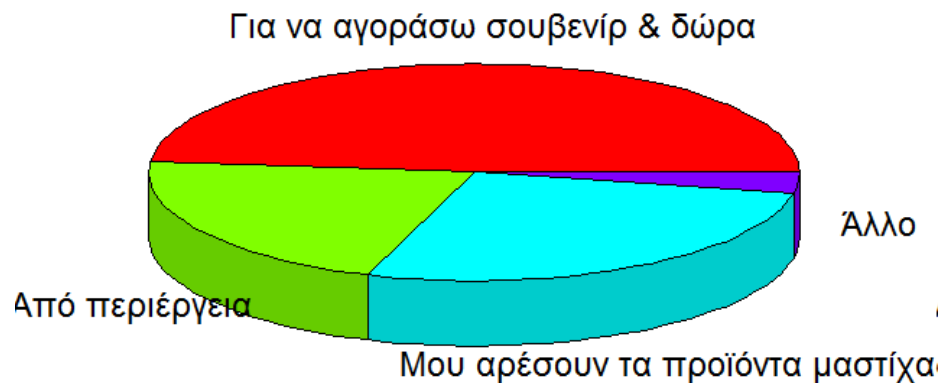


2. Describing Data

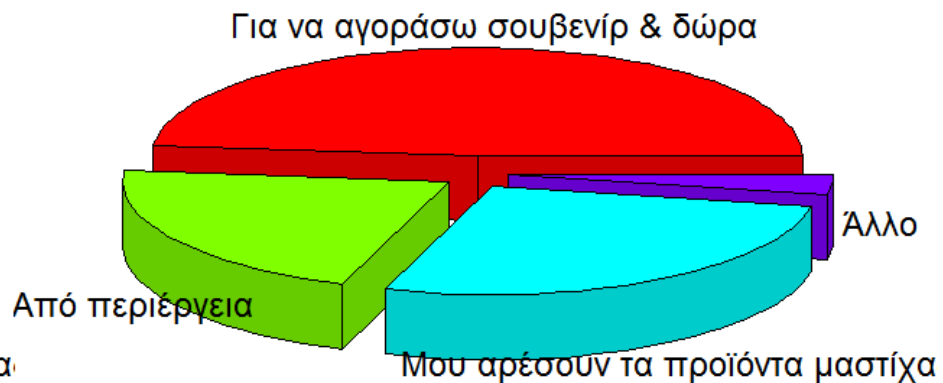
Categorical variables

```
library(plotrix)
slices <- table(masticha$reason.of.visit)
par(mfrow=c(1,2))
pie3D(slices,explode=0, main="Reason of visit")
pie3D(slices,explode=0.1, main="Reason of visit")
```

Reason of visit



Reason of visit

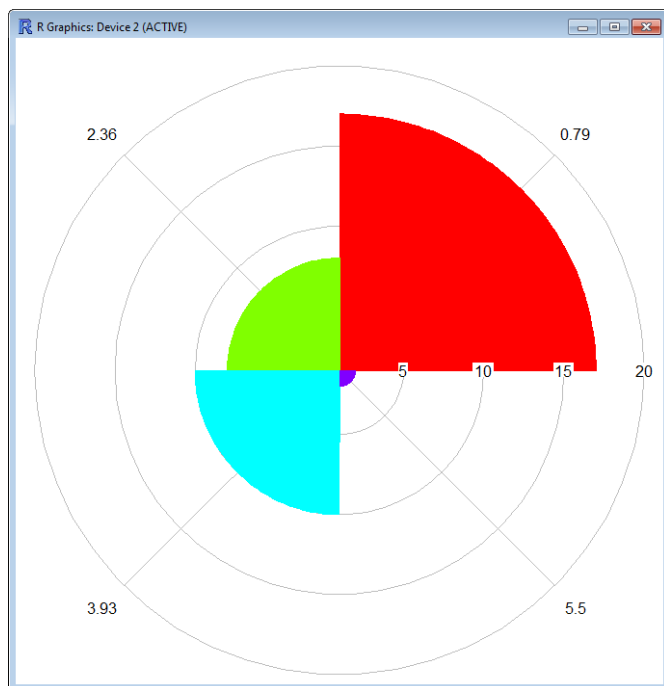


2. Describing Data

Categorical variables



```
library(plotrix)  
slices <- table(masticha$reason.of.visit)  
radial.pie(slices)
```

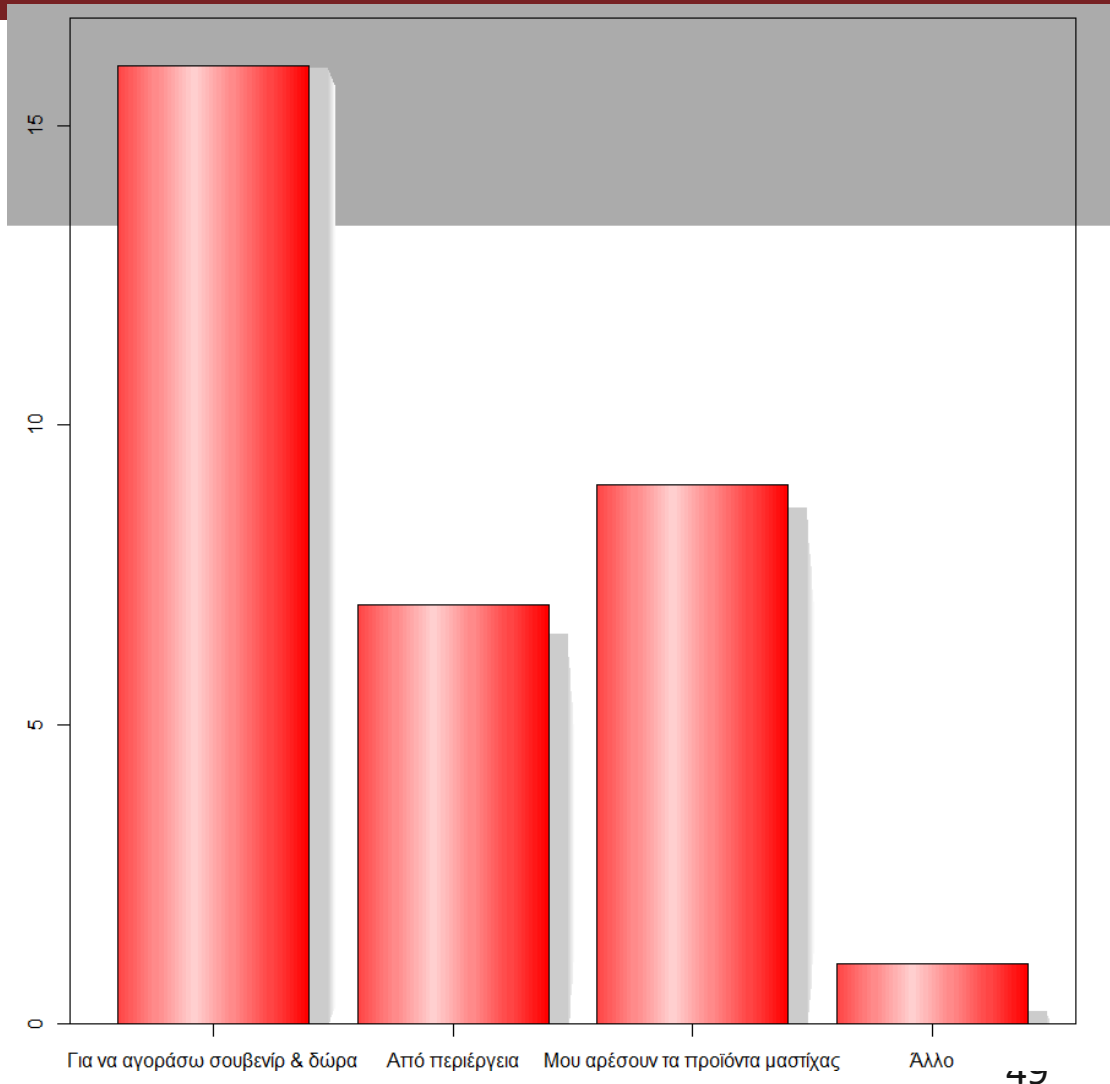


2. Describing Data

Categorical variables



```
library(plotrix)  
x <- table(masticha$reason.of.visit)  
barp(x,col=2, cylindrical=T,  
shadow=T, names.arg=names(x))
```



2. Describing Data

Ordinal variables



- Use them as nominal variables
 - Frequency tabulation
 - Mode
- Cumulative frequencies are meaningful and useful
- We can use
 - The mean and the median as central location measures
 - The standard deviation as measure of dispersionbut carefully especially in interpretation
- The higher the range, more appropriate the methods for quantitative methods are
- Example: Grading or evaluation (1-5, 0-10, 0-100%) 50

2. Describing Data

Ordinal variables

Masticha shop example

masticha\$d9_prices

<i>value</i>	<i>N</i>	<i>raw %</i>	<i>valid %</i>	<i>cumulative %</i>
Πολύ κακή	0	0.00	0.00	0.00
Κακή	1	2.86	3.23	3.23
Μέτρια	8	22.86	25.81	29.03
Καλή	16	45.71	51.61	80.65
Πολύ καλή	6	17.14	19.35	100.00
missings	4	11.43		

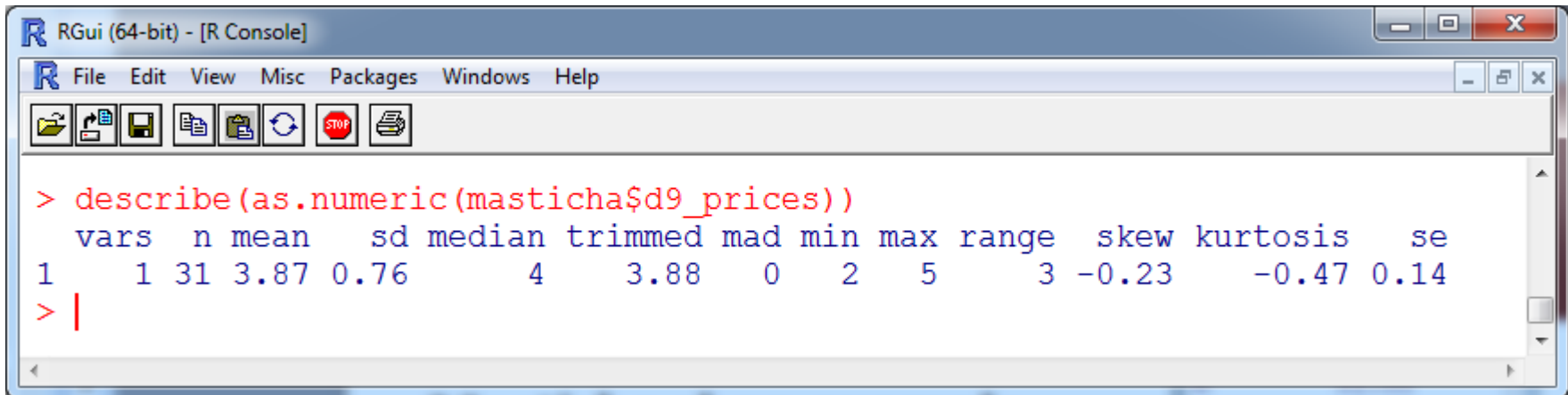
total N=35 · valid N=31 · \bar{x} =3.87 · σ =0.76

2. Describing Data

Ordinal variables

Masticha shop example

```
library(psych)  
describe(as.numeric(masticha$d9_prices))
```



The screenshot shows the RGui (64-bit) [R Console] window. The console displays the following output for the `describe` function:

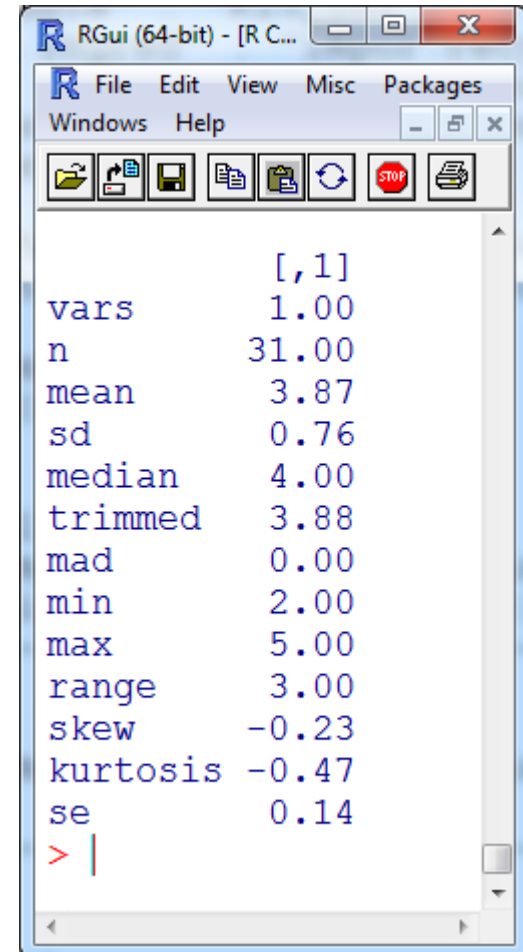
```
> describe(as.numeric(masticha$d9_prices))  
  vars  n mean  sd median trimmed mad min max range  skew kurtosis  se  
1     1 31 3.87 0.76     4   3.88  0  2  5     3 -0.23   -0.47 0.14  
> |
```

2. Describing Data

Ordinal variables

Masticha shop example

```
library(psych)
round(t(describe(as.numeric(masticha$d9
_prices))),2)
```



```
RGui (64-bit) - [R C...
File Edit View Misc Packages
Windows Help
vars      [,1]
vars      1.00
n         31.00
mean      3.87
sd        0.76
median    4.00
trimmed   3.88
mad       0.00
min       2.00
max       5.00
range     3.00
skew      -0.23
kurtosis  -0.47
se        0.14
> |
```