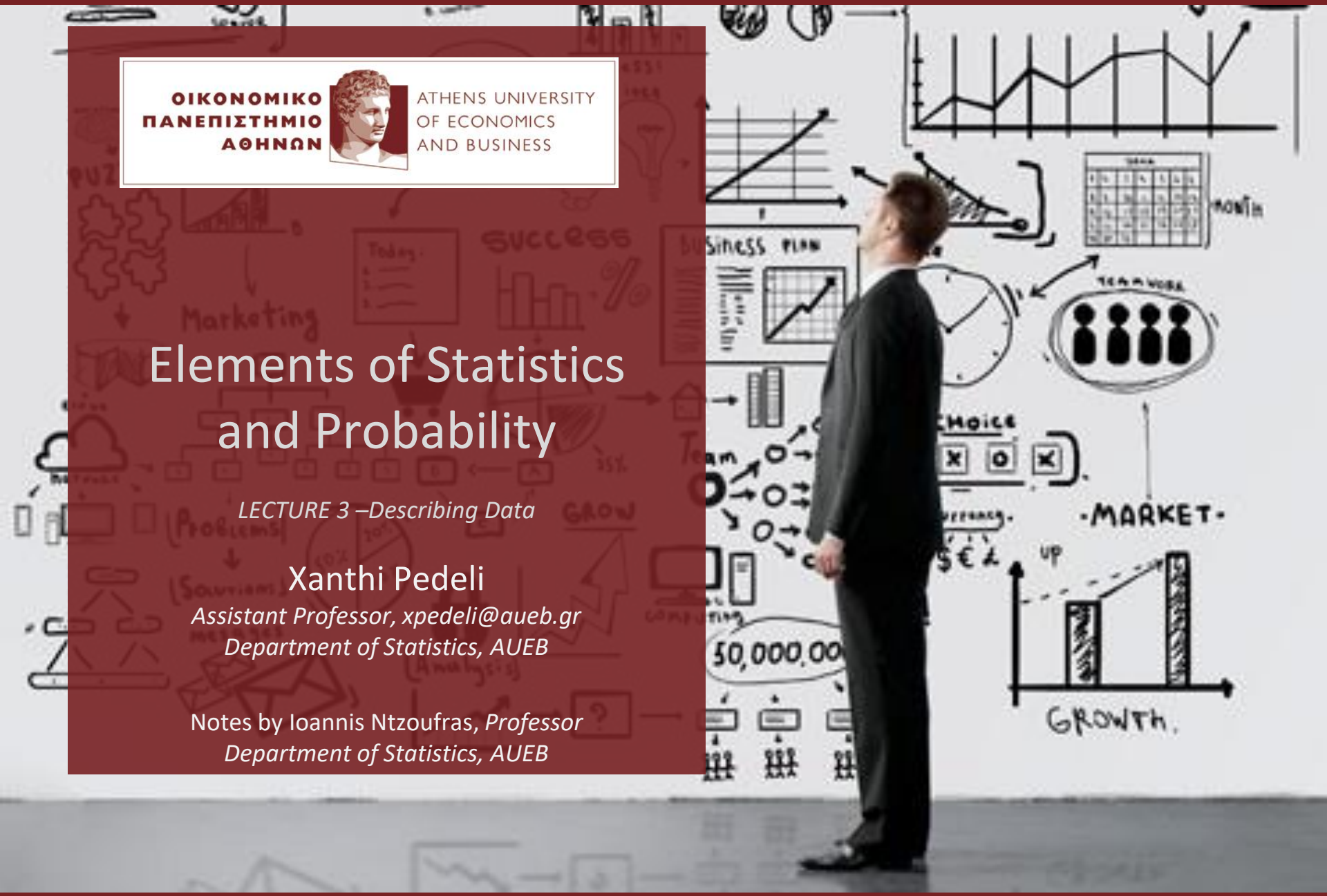


ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

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 ingredientients 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

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:

		Receipt			
	Variable	**1**	**2**	**3**	**4**
1.	**Value (€)**	42	52	48	58
2.	**Book number**	4	5	4	3

2. Describing Data
Organizing data to data frames

ΟΠΑ AUEB

A simple example

- *Observation unit: RECEIPT*
- *Sample size : n=4 receipts*
- *p=2 variables – characteristics:*
  - *Value in Euros*
  - *Number of books*
- *Data matrix*

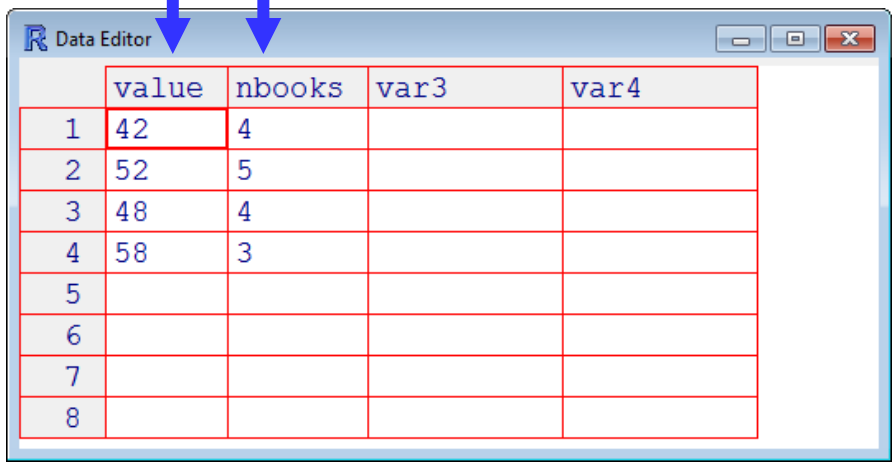
$$\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

2. Describing Data
Organizing data to data frames

A simple example: Specification in R

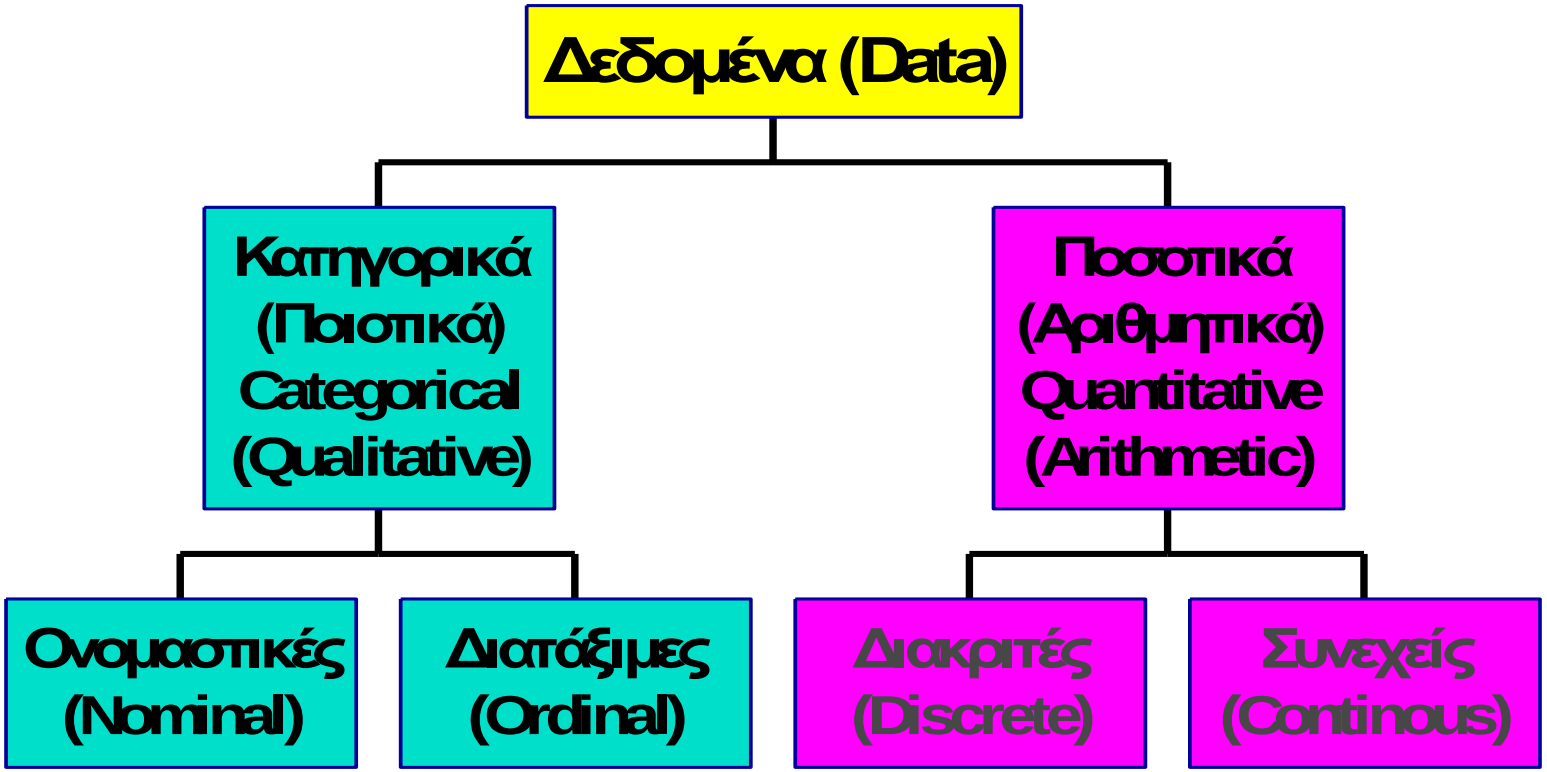
```
ex1 <- data.frame(
          value=c(42,52,48,58),
          nbooks=c(4,5,4,3)  )
ex1<-edit(ex1)
```

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



	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 (kyrtosis, 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

ΟΠΑ AUEB

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/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

ΟΠΑ AUEB

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
  - The 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?

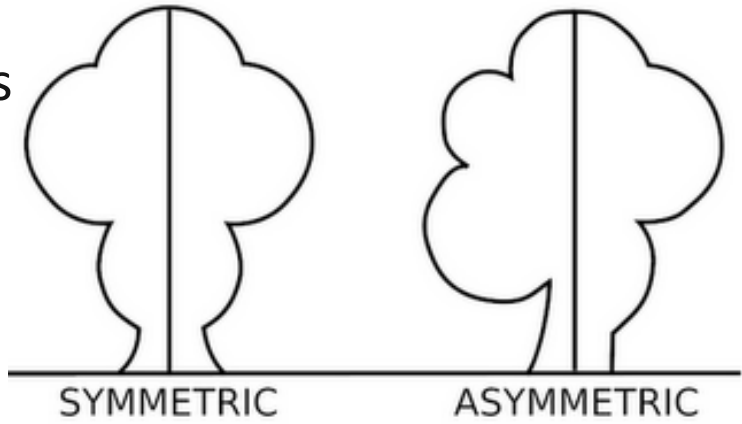
2. Describing Data
Quantitative variables

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

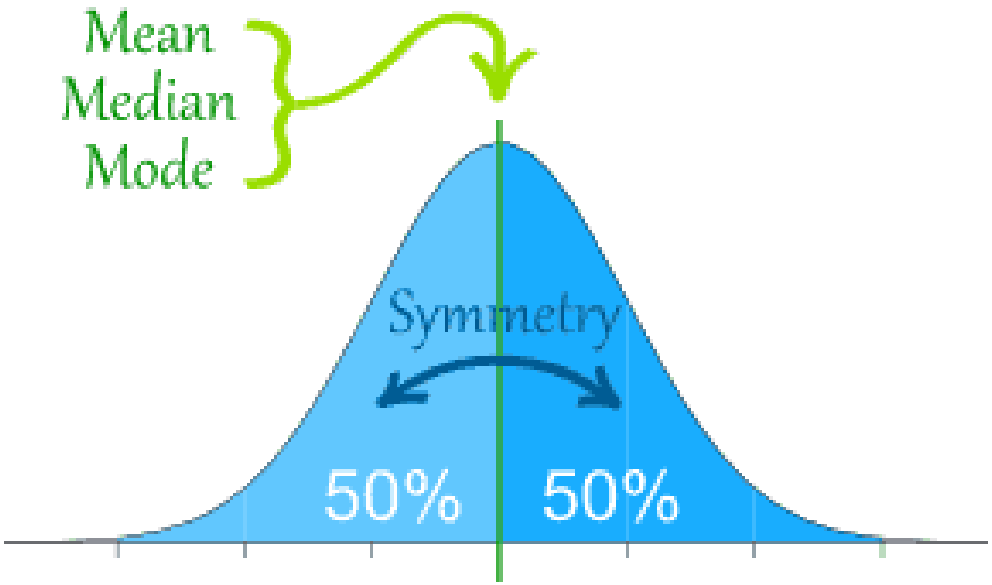
- Measures whether values below and above the mean (probabilistically) behave (appear) in the same way



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

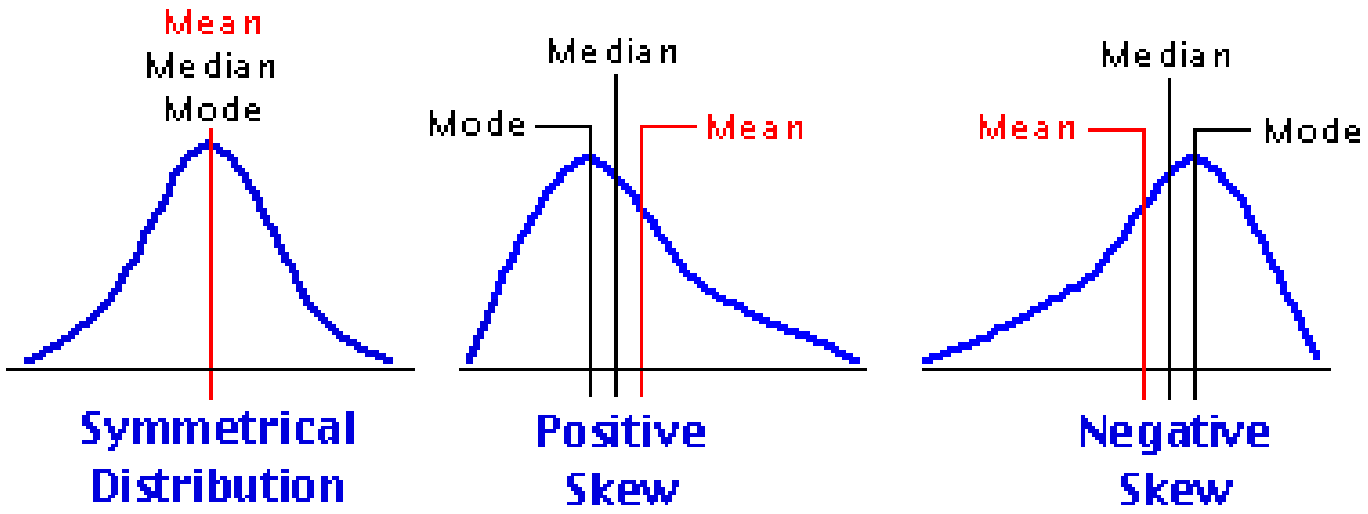


2. Describing Data
Quantitative variables

Symmetry/skewness

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

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

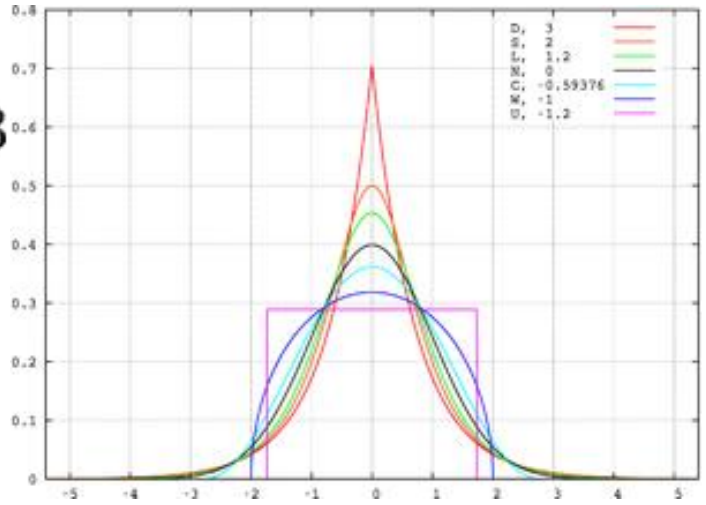


2. Describing Data
Quantitative variables

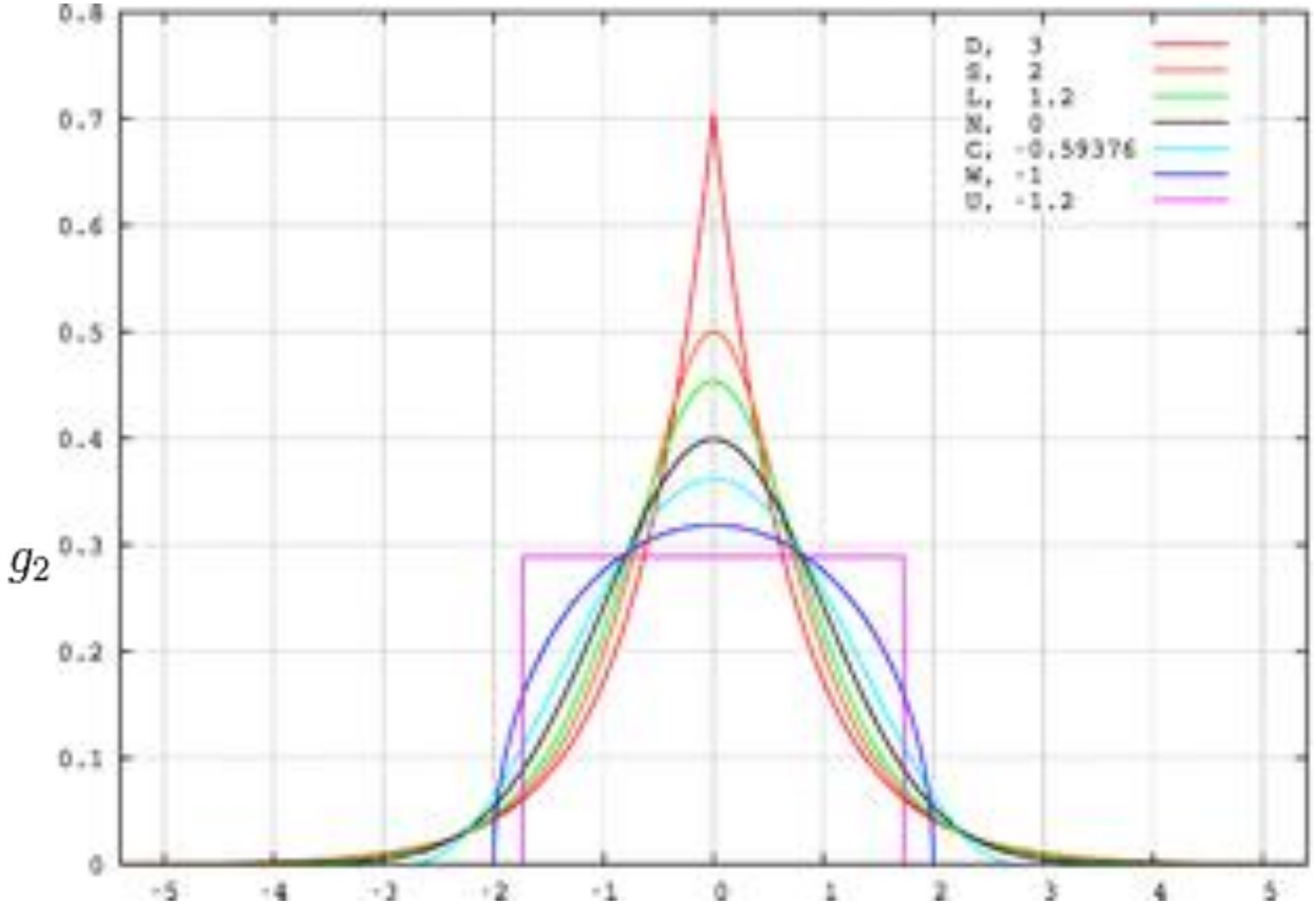
Kutrosis

- 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)

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



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

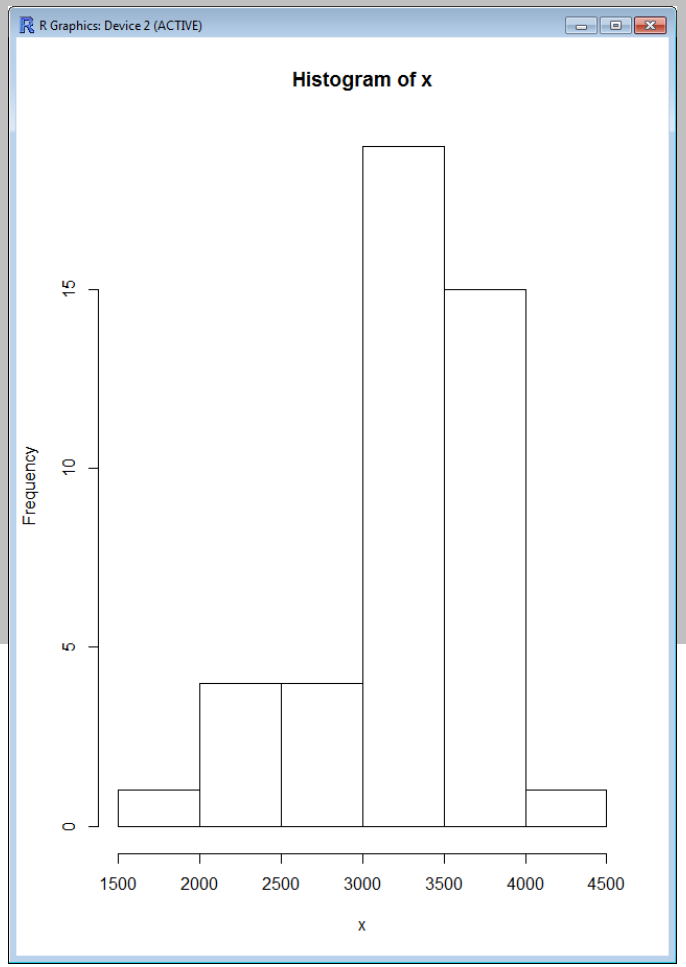
Example – Frequency tabulations

```
> 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

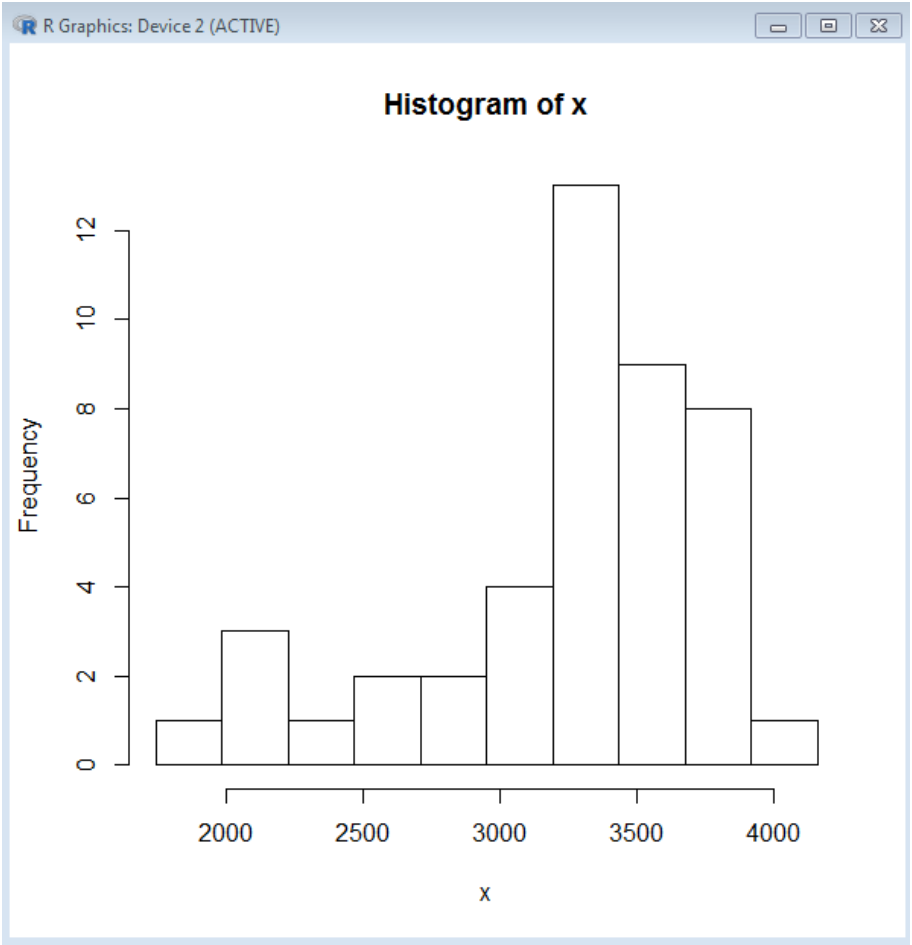
```
factorx <- factor(cut(x, breaks=nclass.scott(x), dig.lab=5))
factorx <- factor(cut(x, breaks=nclass.FD(x), dig.lab=5))

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)
```

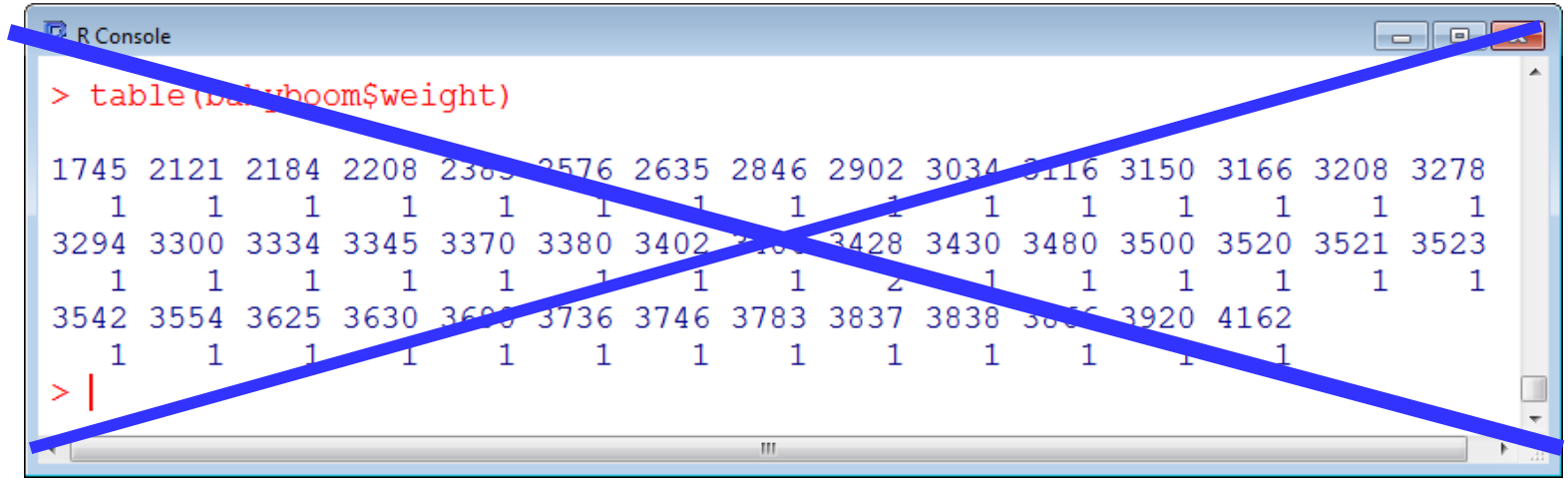
	Freq	cum.Freq	rel.Freq	cum.rel.Freq
[1745,1986.7]	1	1	0.023	0.023
(1986.7,2228.4]	3	4	0.068	0.091
(2228.4,2470.1]	1	5	0.023	0.114
(2470.1,2711.8]	2	7	0.045	0.159
(2711.8,2953.5]	2	9	0.045	0.205
(2953.5,3195.2]	4	13	0.091	0.295
(3195.2,3436.9]	13	26	0.295	0.591
(3436.9,3678.6]	9	35	0.205	0.795
(3678.6,3920.3]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



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

val	*label*	*frq*	*raw.prc*	*valid.prc*	*cum.prc*
[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

val	*label*	*frq*	*raw.prc*	*valid.prc*	*cum.prc*
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

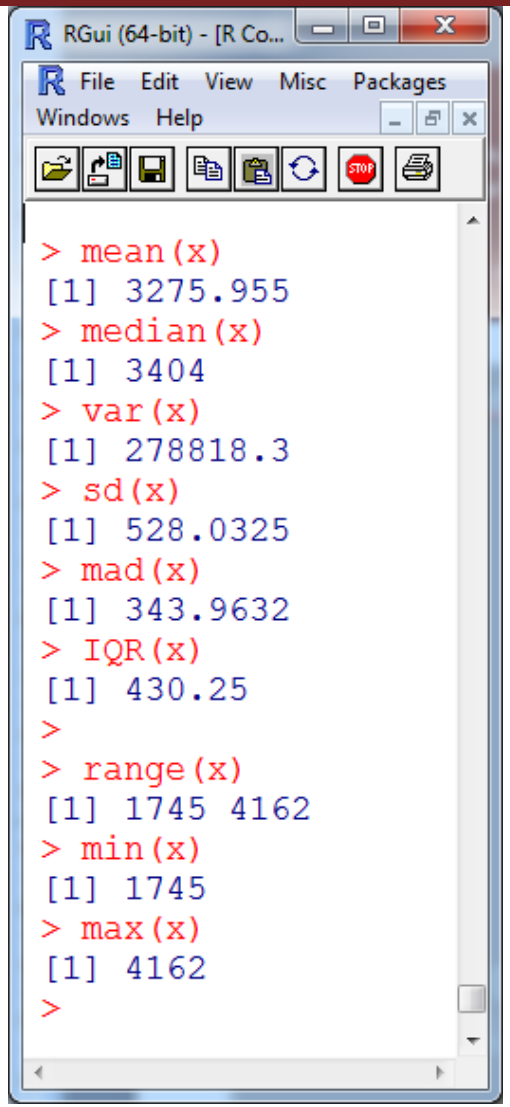
Example – Descriptive measures

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 )
```

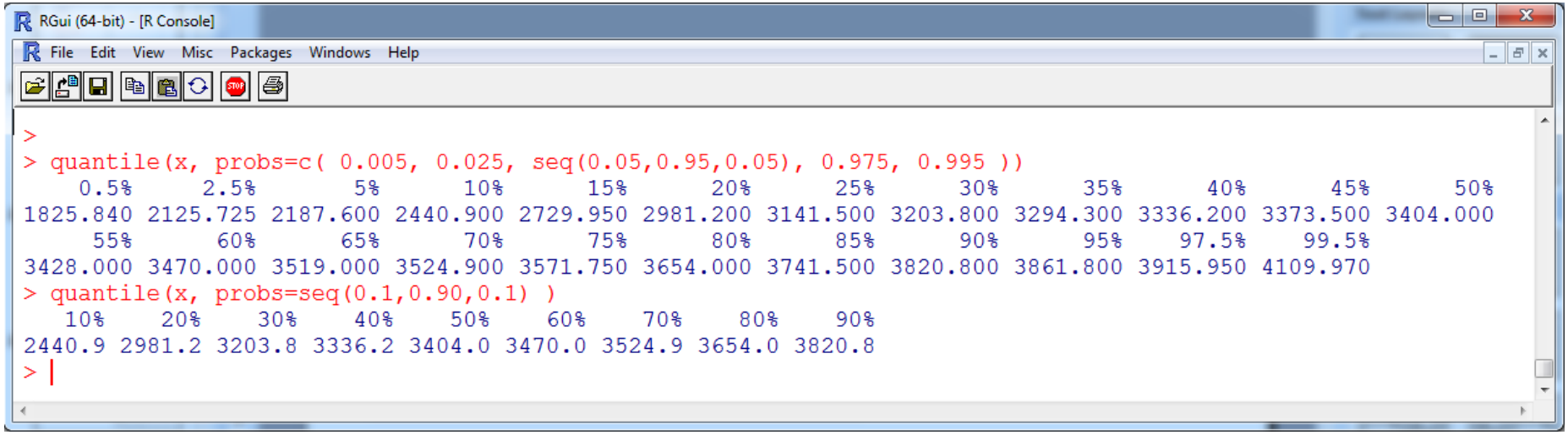


2. Describing Data

Example – Descriptive measures

Quantiles in R using vectors

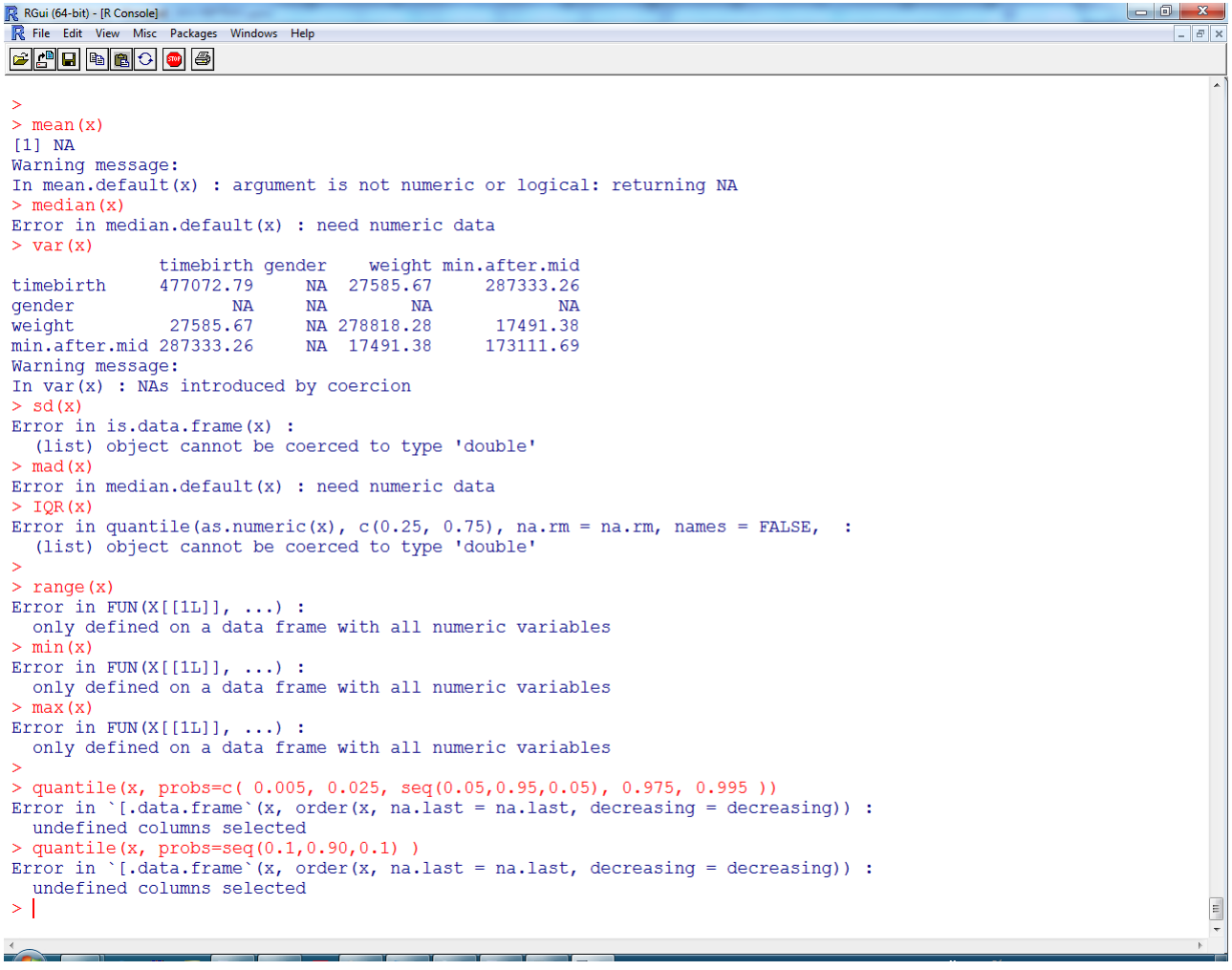
```
quantile(x, probs=pp)
quantile(x, probs=seq(0.1,0.90,0.1) )
```



2. Describing Data

Example – Descriptive measures

Descriptives in R – using data frames



2. Describing Data

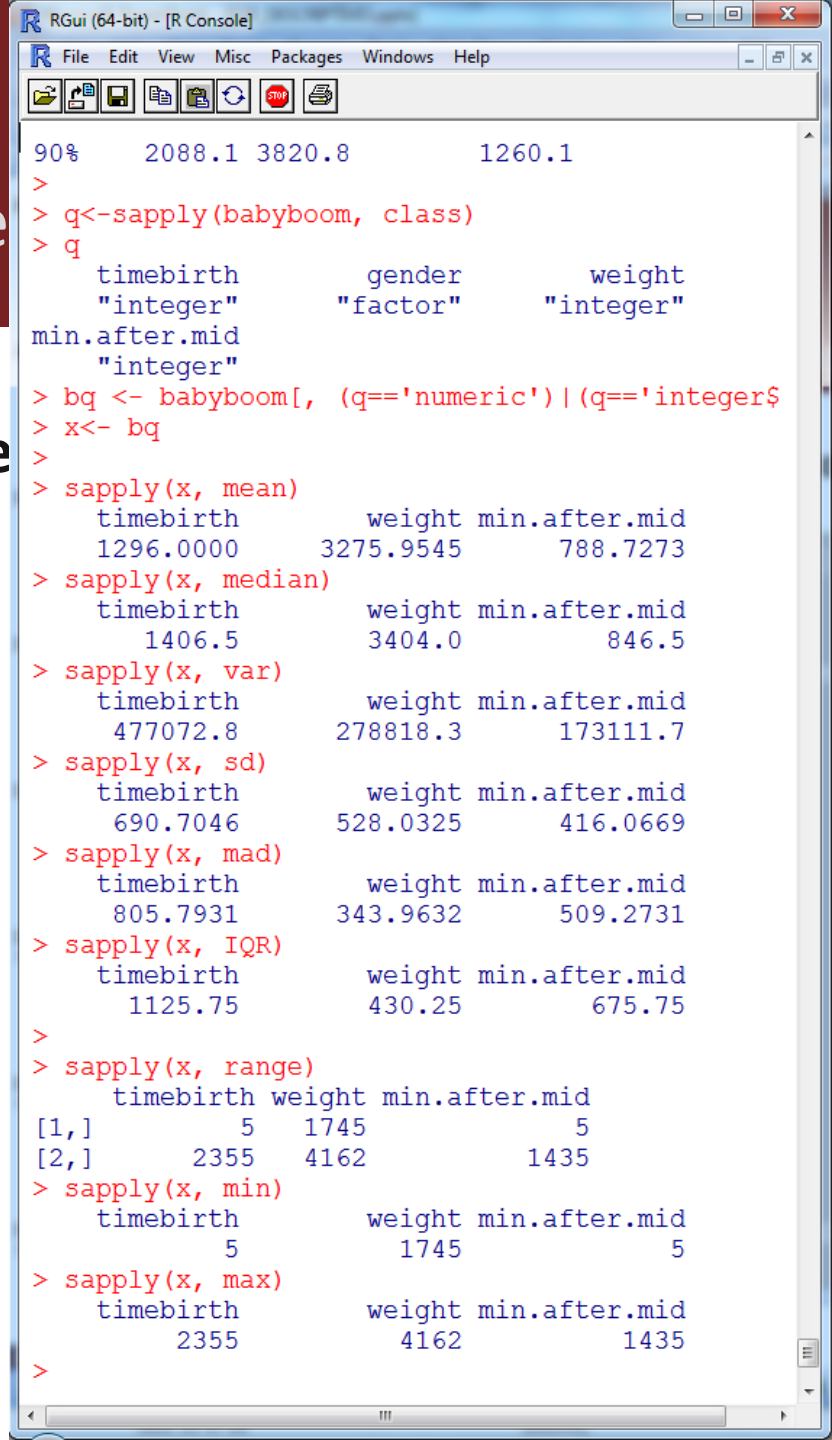
Example – Descriptive measure

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)
```



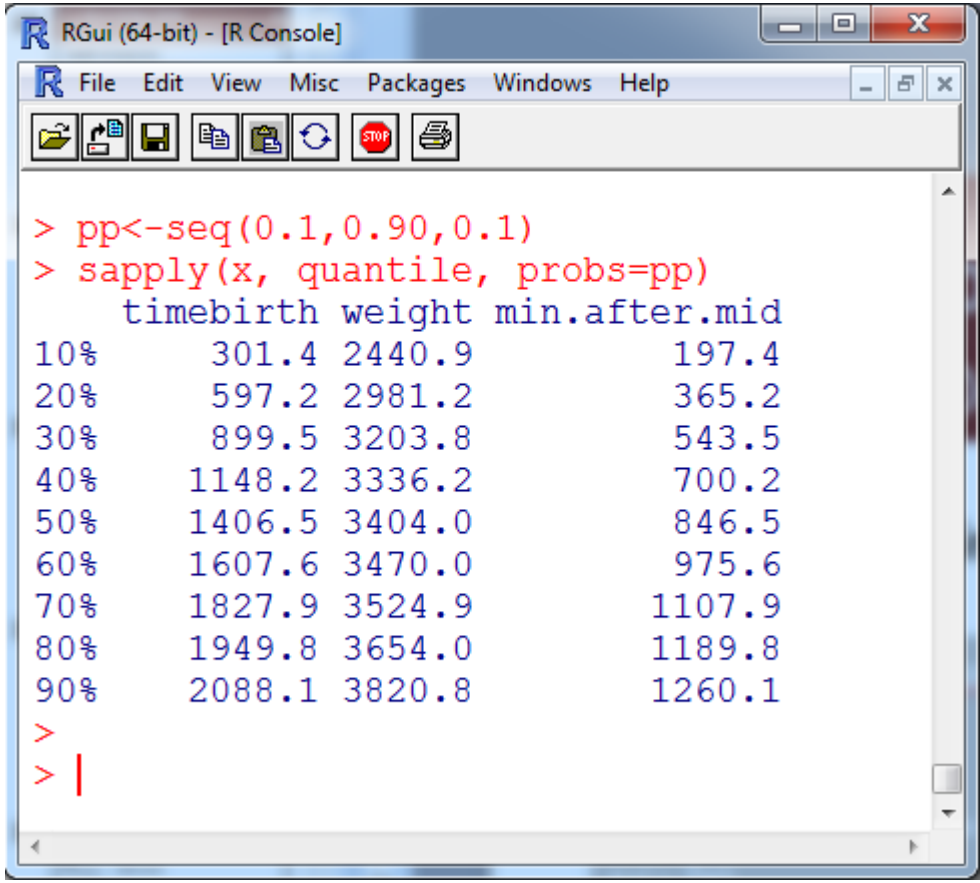
2. Describing Data

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)
```

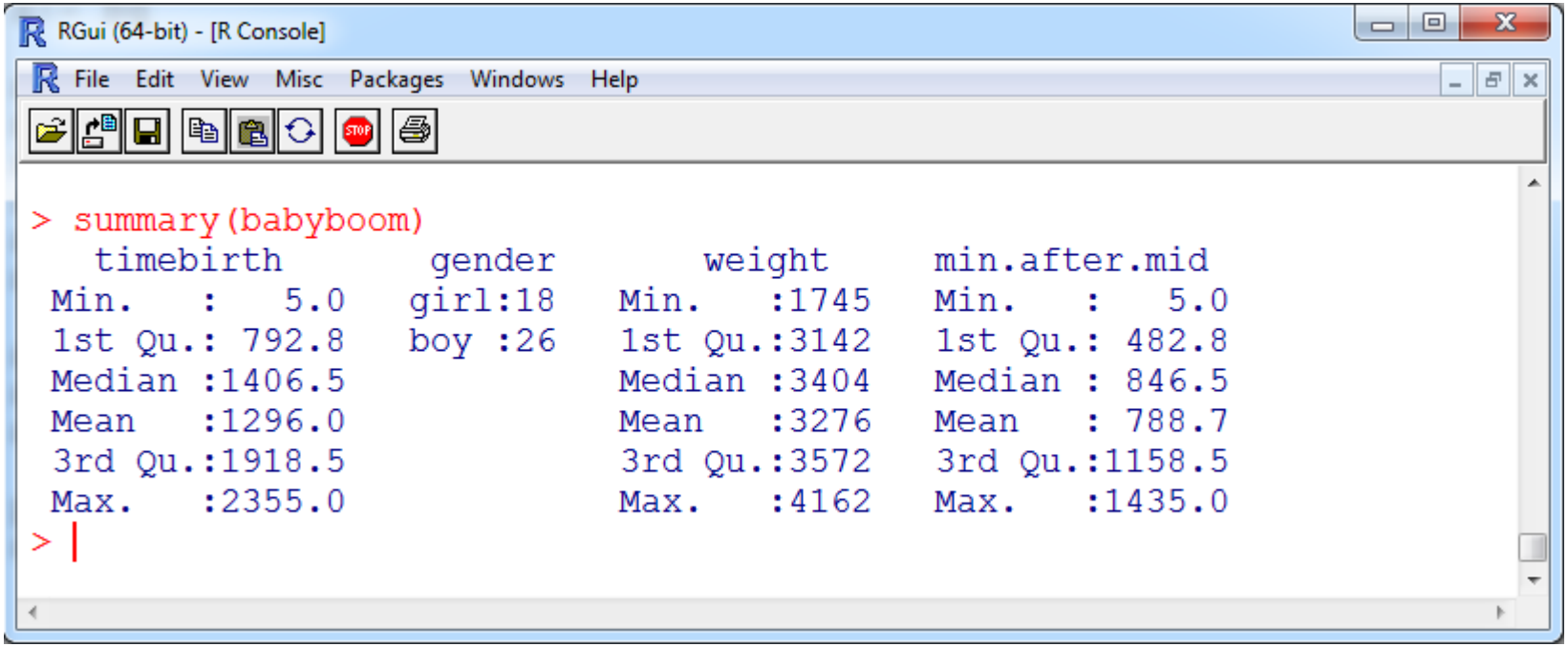


2. Describing Data

Example – Describing measures

Summary statistics of dataframes

```
summary(babyboom)
```

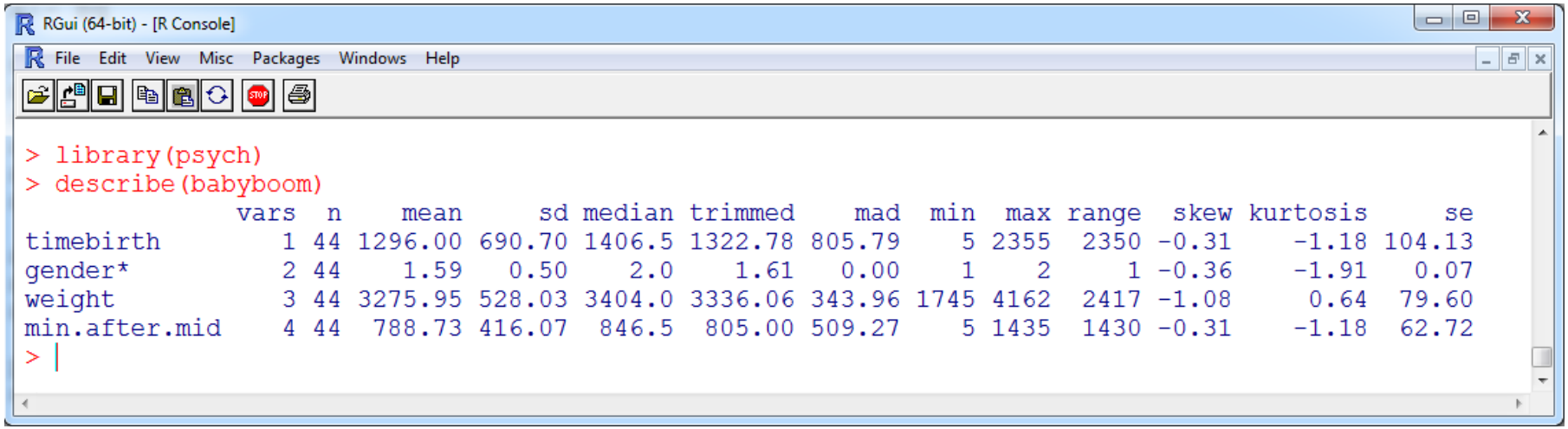


2. Describing Data

Example – Descriptive measures

Summary statistics of dataframes

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

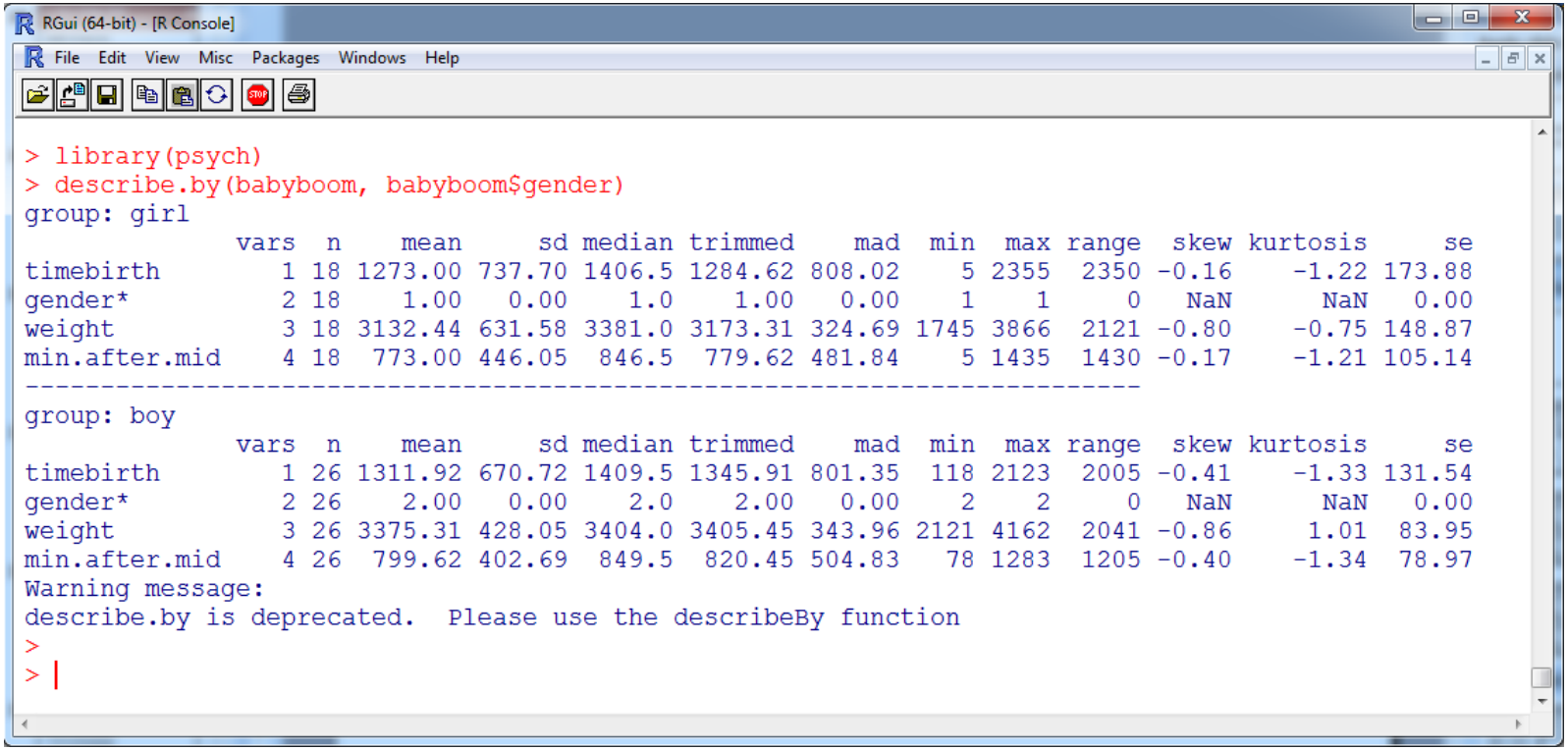


2. Describing Data

Example – Descriptive measures

Summary statistics of dataframes

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



Categorical variables – nominal or qualitative

- Frequency tabulation
- The mode

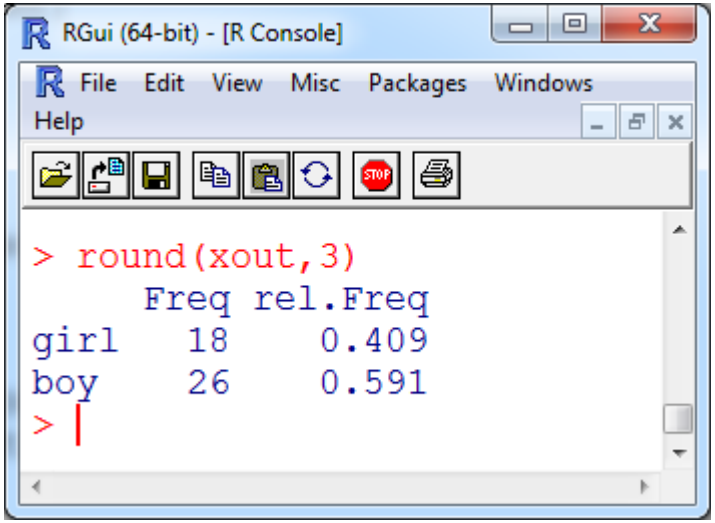
2. Describing Data
Categorical variables

ΟΠΑ AUEB

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)
```



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

val	*l*
girl	
boy	
NA	N

Gender of newborn

val	*label*	*frq*	*raw.prc*	*valid.prc*	*cum.prc*
girl		18	40.91	40.91	40.91
boy		26	59.09	59.09	100
NA	NA	0	0	NA	NA

total N=44 · valid N=44 · x̄=1.59 · σ=0.50

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

value	*N*	*raw %*	*valid %*	*cumulative %*
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

value	*N*	*raw %*	*valid %*	*cumulative %*
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

value	*N*	*raw %*	*valid %*	*cumulative %*
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			
		tota		

R Console

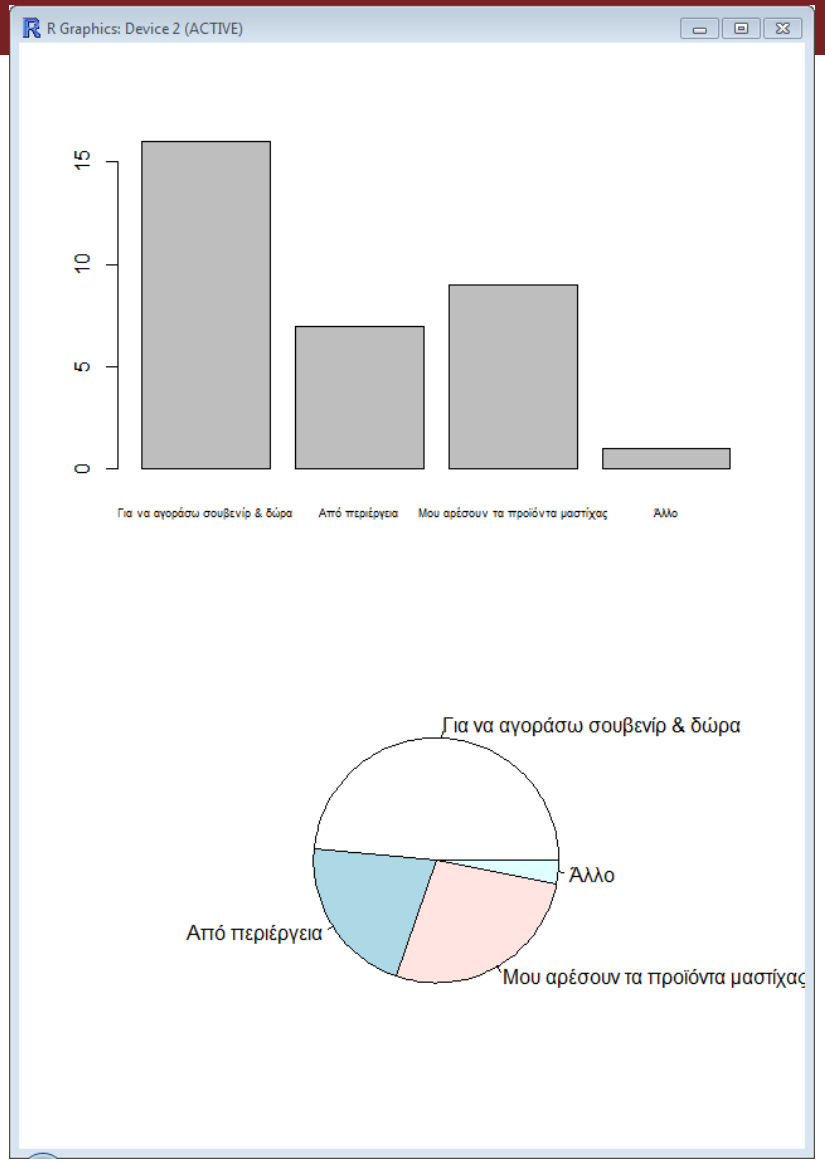
```
> tabx <- table(x)
> tabx
x
  Για να αγοράσω σουβενίρ & δώρα                 Από περιέργεια
                              16                              7
Μου αρέσουν τα προϊόντα μαστίχας                          Άλλο
                               9                              1
> tabx[which(tabx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

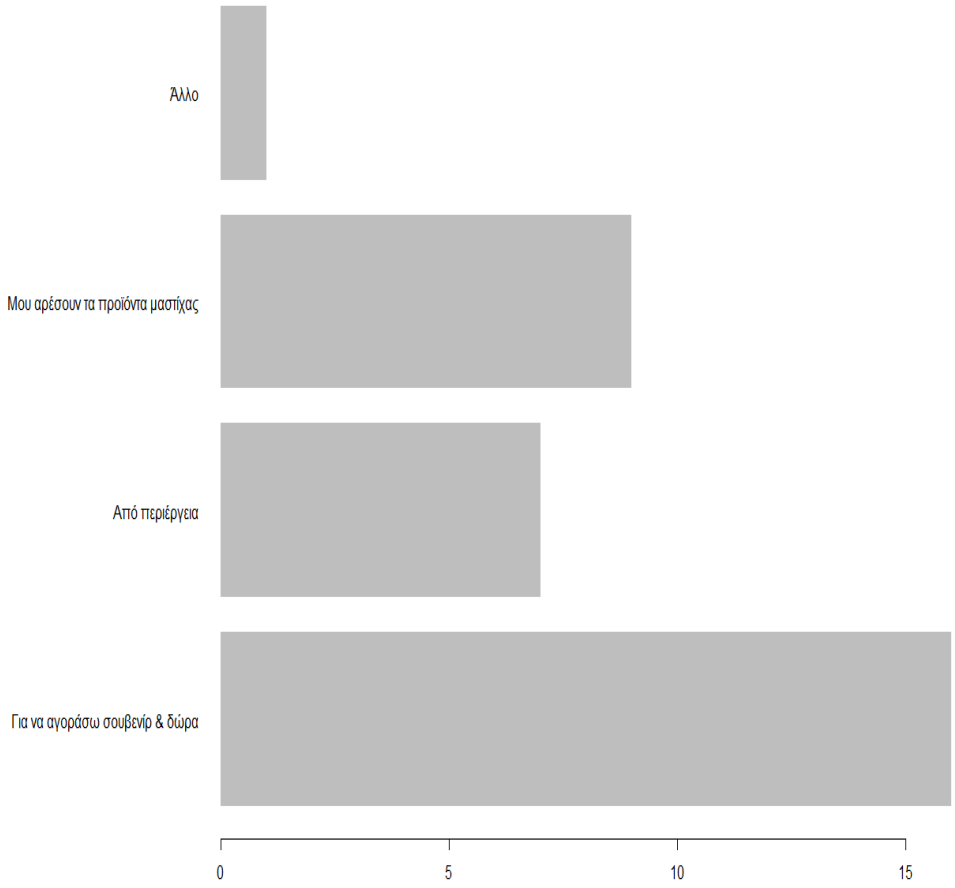
```
par(mfrow=c(2,1))
barplot(table(masticha$reason.of.visit))
pie(table(masticha$reason.of.visit))
```



2. Describing Data
Categorical variables

Masticha Shop dataset

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

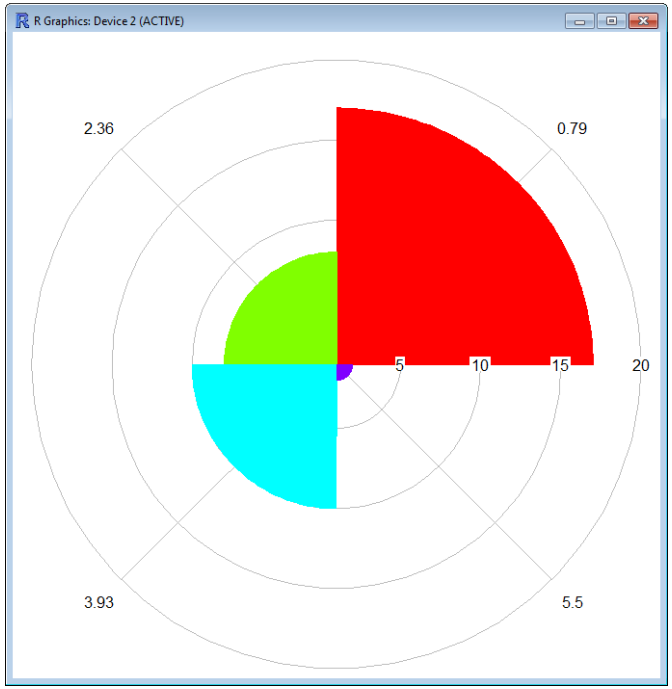


2. Describing Data
Categorical variables

```
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")
```

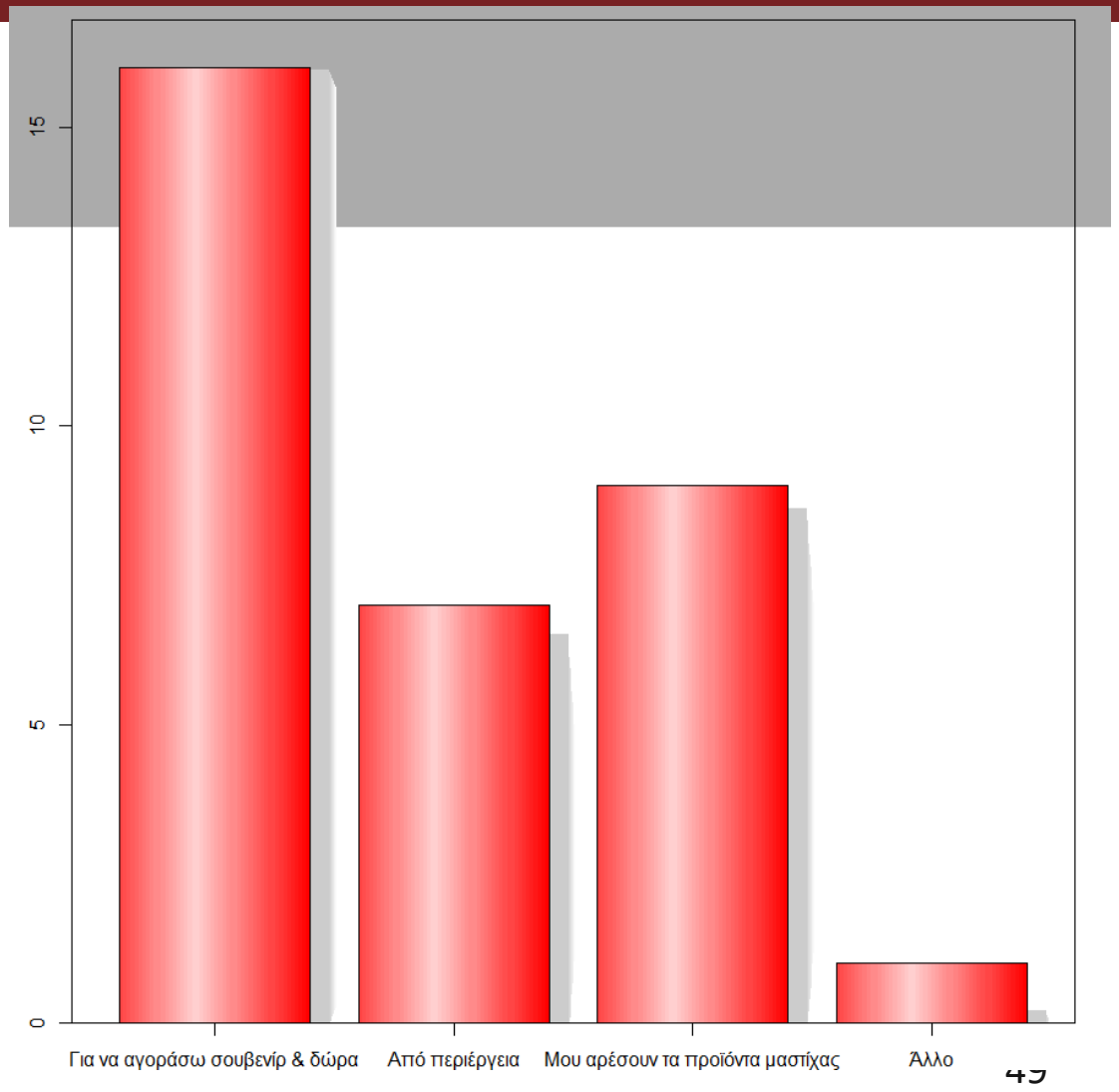
2. Describing Data
Categorical variables

```
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 dispersion

  but 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%)

2. Describing Data
Ordinal variables

Masticha shop example

masticha$d9_prices

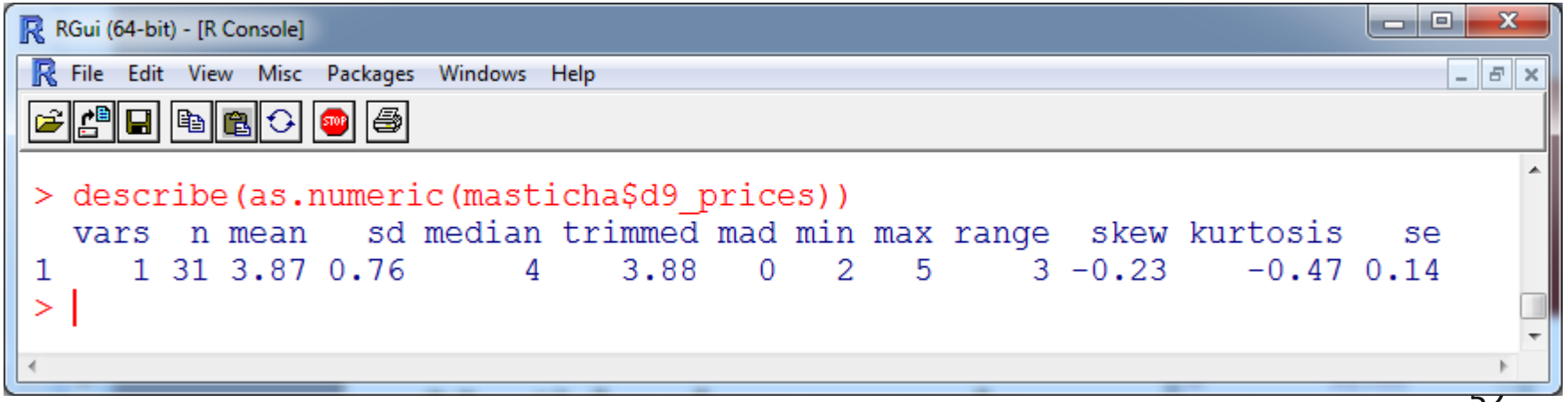
value	*N*	*raw %*	*valid %*	*cumulative %*
Πολύ κακή	0	0.00	0.00	0.00
Κακή	1	2.86	3.23	3.23
Μέτρια	8	22.86	25.81	29.03
Καλή	16	45.71	51.61	80.65
Πολύ καλή	6	17.14	19.35	100.00
missings	4	11.43		

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))
```



2. Describing Data
Ordinal variables

Masticha shop example

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