contents

Outillic

Histograms

KDES

Scatterplot

_ .

Болріось

Massis slat

QQ-plots

Heatmans

Lecture 11

Ciprian M. Crainiceanu

Department of Biostatistics Johns Hopkins Bloomberg School of Public Health Johns Hopkins University

August 21, 2021

Table of contents

Table of contents

. . .

...---

C - - 11 - - - - 1 - 1

Boxplots

Mosaic plot

. . . .

Heatmap

- 1 Table of contents
- 2 Outline
- 3 Histograms
- 4 KDEs
- Scatterplots
- **6** Dotcharts
- Boxplots
- 8 Bar plots
- Mosaic plots
- QQ-plots
- Heatmaps

Table of

Outline

Histogram

ocatter prots

Boynlo

Bar pla

Mosaic plot

QQ-plots

Heatman

Outline

- ① Histograms
- 2 Stem-and-leaf plots
- 3 Dot charts and dot plots
- 4 Boxplots
- **5** Kernel density estimates
- QQ-plots

Table of contents

Outline

Histograms

KDE

Scatterplot

Dottilaits

Boxplots

QQ-plot

Heatmaps

Histograms

 Histograms display a sample estimate of the density or mass function by plotting a bar graph of the frequency or proportion of times that a variable takes specific values, or a range of values for continuous data, within a sample

Table of content

Histograms

KDEs

Scatterplot

Dotcharte

Boxplot

Dui pioto

...osaic pic

d d piots

Example: Frequency histograms

- Sleep Heart Health Data
- WASO (Wake After Sleep Onset): linked to sleep quality.
- Expressed in minutes

Table of contents

0 44.....

Histograms

KDE:

Scatterplo

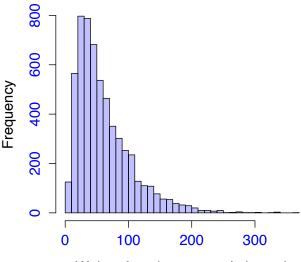
_ .

2

Mosaic plot

. . . .

Histogram of frequencies



Wake after sleep onset (minutes)

Table of

Outline

Histograms

KDE:

Scatterplot

Boxplot

.

00 1.

~ · · · · · ·

Example: Probability distribution histograms

Table of contents

Histograms

KDEs

Scatterplo

Dottina

Doxpic

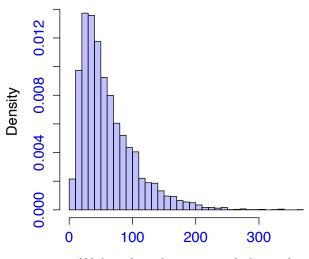
Dai piot

Mosaic plot

QQ-plots

Heatman

Histogram of probability distribution



Wake after sleep onset (minutes)

Table of contents

Histograms

Scatterplot

D

. .

Mosaic plot

QQ-plots

Heatmap

Pros and cons

- Histograms are useful and easy, apply to continuous, discrete and even unordered data
- They use a lot of ink and space to display information
- It is difficult to display several at the same time
- Certain distributions may require data transformation for proper plotting

Table of contents

. . .

listogram

KDEs

Scatterplot

ь .

Manadamilan

. . . .

Kernel density estimates

- Kernel density estimates are essentially more modern versions of histograms providing density estimates for continuous data
- Observations are weighted according to a "kernel", in most cases a Gaussian density
- "Bandwidth" of the kernel effectively plays the role of the bin size for the histogram
 - Too low of a bandwidth yields a too variable (jagged) measure of the density
 - **b** Too high of a bandwidth oversmooths
- The R function density can be used to create KDEs

Table of

0 4 4 1 1 1 1

Histograms

KDEs

Scatterplot

Daniel

Bar plo

Mosaic plot

QQ-plots

Heatmaps

Example: Automatic KS

Table of

Outline

Histograr

KDEs

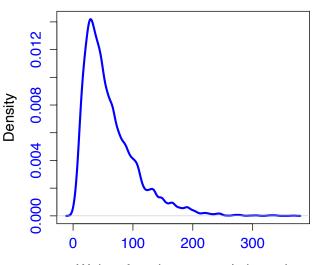
Scatterplo

_ .

QQ-plots

Heatman

Kernel density estimate



Wake after sleep onset (minutes)

Table of

Histograms

KDEs

Scatterplot

Boxplots

Manaia alah

QQ-plots

Heatman

Example: Automatic KS

Table of

Outille

Histogram

KDEs

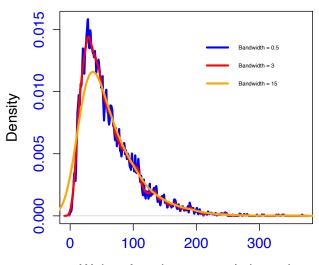
Scatterplot

Boynlo

QQ-plots

Heatman

Kernel density estimate (Gaussian)



Wake after sleep onset (minutes)

Example

Table of

listogram:

KDEs

ocatte piot

Roynlots

_

Mosaic plot

QQ-plots

Heatmap

Data is the waiting and eruption times in minutes between eruptions of the Old Faithful Geyser in Yellowstone National park

```
data(faithful)
d <- density(faithful$eruptions, bw = "sj")
plot(d)</pre>
```

Ciprian M.

Table of

Outline

Histogram

KDE-

Scatterplot

Database

Boxplo

Bar plo

Mosaic plots

OO ploto

Heatman

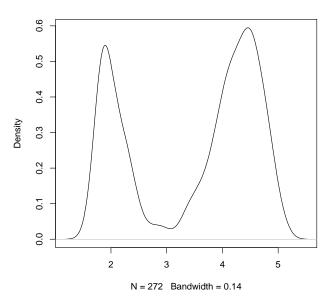


Table of

Outline

Histogram

KDEs

Scatterplot

Manager of the

00 -1-+-

Heatmap:

Imaging example

- Consider the following image slice (created in R) from a high resolution MRI of a brain
- This is a single (axial) slice of a three-dimensional image
- Consider discarding the location information and plotting a KDE of the intensities

Ciprian M.

Table of

Outlin

distograme

KDEs

Scatterplot

Dotcharts

Povolet

Bar plot

Mosaic plots

00 1.

Heatmans

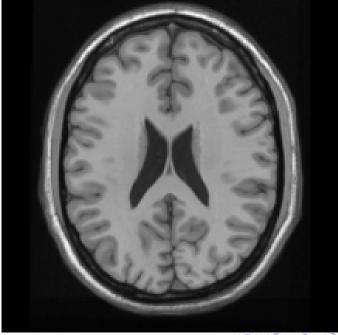


Table of contents

Outline

Histogra

KDEs

Scatterplot

Danielat.

Болріос.

...----

QQ-plots

Heatmans

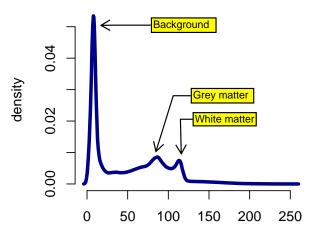


Table of contents

Outline

Histogram:

KDE

Scatterplots

Boxplots

Par plate

Mosaic plots

OO-plots

Heatmaps

Scatterplots

- Histograms and KDEs: display marginal distributions
- Scatterplots: display the joint distribution of two variables
- Marginal and 2D scatterplots are great for initial data exploration
- BMI versus RDI 4%, with lowess and mgcv smoothers

Table of contents

Outillic

Histogram

Scatterplots

Boxplot

_

QQ-plots

Heatmap

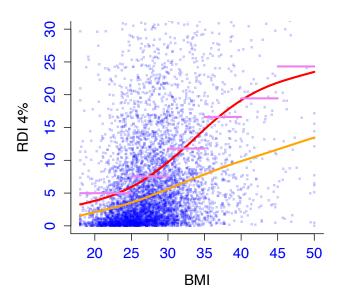


Table of contents

Thistogram

Scattorplat

ocatte prot

Dotcharts Boxplots

Bar plots

Mosaic plot

QQ-plots

неаттар

Dotcharts

- Dotcharts simply display a data set, one point per dot
- Ordering of the dots and labeling of the axes can display additional information
- Dotcharts show a complete data set and so have high information density
- May be impossible to construct/difficult to interpret for data sets with lots of points

library(datasets)
dotchart(log10(islands))

Victoria

imor

aiwan Sumatra

akhalin

loluccas

Melville

uzon Kvushu reland celand Jonshu

Haihan

Europe

Devon Cuba Celebes Britain Borneo

Asia

Africa

Dotcharts

islands data: log10(area) (log10(sq. miles))

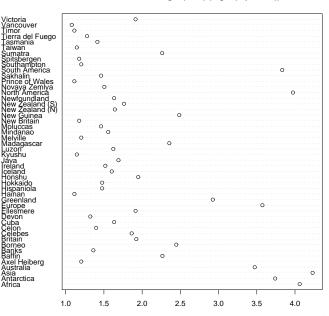


Table of

Outline

listograms

KDE:

Scatterplot

Dotcharts

D

_ ----

iviosaic pio

QQ-plot

Heatmaps

Discussion

- Maybe ordering alphabetically is not the best thing for this data set
- Perhaps grouped by continent, then nations by geography (grouping Pacific islands together)?

Table of contents

Outline

Histograms

NDES

Scatterplots

Dotcharts

Boxplots

Rar plots

Mosaic plot

QQ-plots

Heatman

Dotplots comparing grouped data

- For data sets in groups, you often want to display density information by group
- If the size of the data allows it, displaying the whole data is preferable
- Add horizontal lines to depict means, medians
- Add vertical lines to depict variation, show confidence intervals interquartile ranges
- Jitter the points to avoid overplotting (jitter)

0 4 4 1 1 1

ilistograffis

KDES

Scatterplot

Dotcharts

Boxplots

Mosaic plot

QQ-plot

Heatmaps

- The InsectSprays dataset contains counts of insect deaths by insecticide type (A, B, C, D, E, F)
- You can obtain the data set with the command data(InsectSprays)

Table of contents

nistograms

Scatterplot

Dotcharts

Бохрюсѕ

Mosaic plot

00 -1-+-

Heatman

The gist of the code is below

```
attach(InsectSprays)
plot(c(.5, 6.5), range(count))
sprayTypes <- unique(spray)</pre>
for (i in 1 : length(sprayTypes)){
  y <- count[spray == sprayTypes[i]]</pre>
  n <- sum(spray == sprayTypes[i])</pre>
  points(jitter(rep(i, n), amount = .1), y)
  lines(i + c(.12, .28), rep(mean(y), 2), lwd = 3)
  lines(rep(i + .2, 2),
        mean(y) + c(-1.96, 1.96) * sd(y) / sqrt(n)
```

Ciprian M.

Table of

Outillic

Histogram

KDE

Dotcharts

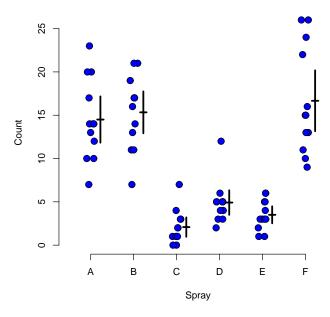
Royalot

Болріос

...----

QQ-plots

Haatman



Outline

stogram

KDEs

Scatterplots

Boxplots

Бохрюсѕ

iviosaic piot

QQ-plots

Heatmaı

- Boxplots: when displaying every point is not possible
- Centerline is the median; the box edges are the quartiles
- Whiskers: a constant times the IQR or the max value
- Sometimes outliers are points beyond the whiskers
- Also invented by Tukey
- Skewness indicated by centerline being near one of the box edges

Table of contents

....

i iistograiii:

c

ocatter prot

Boxplots

Болріось

. . . .

00

Heatman

Example: SHHS

- Distribution of RDI 4% in two groups.
- Group 1: BMI<25 and age<50
- Group 2: BMI>35 and age>70

Ciprian M.

Table of

0 4 4 1 1 1

Histograms

KDEs

Scatterplot

Dotcharts

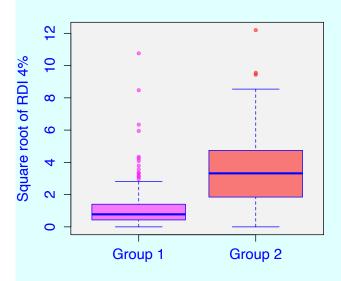
Boxplots

Болріось

iviosaic piot

QQ-plots

Heatman



Example: SHHS

Table of contents

Outille

nistogram

INDES

Scatterplot

 ${\sf Boxplots}$

ь .

Mosaic plot

00 -1-+-

Heatman

Distribution of RDI 4% in six BMI groups.

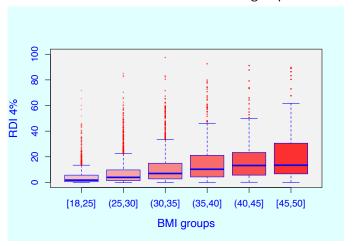


Table of contents

Outilite

Histograms

INDLS

Scatterpiot

Boxplots

Dui piots

Mosaic plo

QQ-plots

Heatman

Example: SHHS

- Distribution of RDI 4% in the same six BMI groups.
- Now by females/males

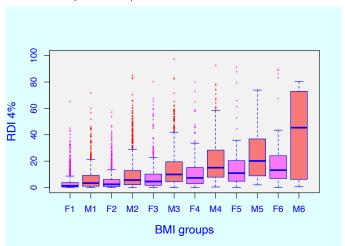


Table of contents

Outline

Histograms

.....

Scatterplots

Boxplots

Doxplots

Mosaic plot

QQ-plots

Heatmap

Boxplots discussion

- Don't use boxplots for small numbers of observations, just plot the data!
- Try logging if some of the boxes are too squished relative to the other ones; you can convert the axis to unlogged units (though they will not be equally spaced anymore)
- For data with lots and lots of observations omit the outliers plotting if you get so many of them that you cant see the points

. . .

Scattorplot

Boxplo

Bar plots

Mosaic plots

QQ-plots

Heatmaps

- Most useful for indicating number of observations in groups
- There are different types of bar plots
- Useful in many publications

Table of

Histograms

INDLS

Scatterplot

Dotcharts

Bar plots

Mosaic plot

QQ-plots

Heatmaps

Data structure for a simple bar plot

```
counts <- table(bmi_cut)</pre>
```

> counts

bmi_cut

```
[18,25] (25,30] (30,35] (35,40] (40,45] [45,50]
1608 2428 1190 371 116 48
```

Table of contents

...---

ocatter prote

D

Bar plots

Manageria

iviosaic piot.

QQ-plots

Heatmap

Code for a simple bar plot

```
par(bg="lightcyan")
plot(3, 3, type="n", ann=FALSE, axes=FALSE)
# The coordinates of the plot area
u <- par("usr")</pre>
rect(u[1],u[3],u[2],u[4], col="gray95", border=NA)
par(new=TRUE)
barplot(counts, main="",
    vlab="Number of individuals",
    xlab="BMI groups", col=rgb(1,0,0,0.6),
    vlim=c(0,2500),cex.axis=1.3,
    col.axis="blue",cex.lab=1.3,
    col.lab="blue")
```

Ciprian M.

Table of

Outilite

Histogram

INDLS

Scatterploi

Daniel

Bar plots

Mosaic plots

OO plots

Heatmap

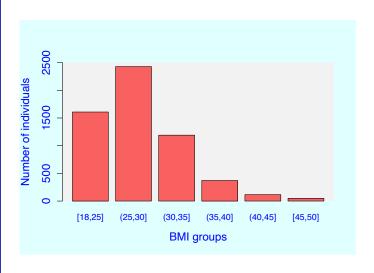


Table of contents

0 4

nistograms

KDE:

Scatterpio

Boxplo

Bar plots

Mosaic plots

QQ-plots

Heatmap

Data structure for a stacked bar plot

counts <- table(smokstatus,bmi_cut)</pre>

> counts

bmi_cut

smokstatus	[18,25]	(25,30]	(30,35]	(35,40]	(40,45]	[45,50]
Never	793	1118	541	165	49	22
Current	206	222	95	17	10	5
Former	603	1073	545	185	56	21

Table of

Outline

∐icto aromo

KDE

Scatterplo[®]

Royale

Bar plots

Mosaic plots

QQ-plots

Heatmap

Boxplot of counts

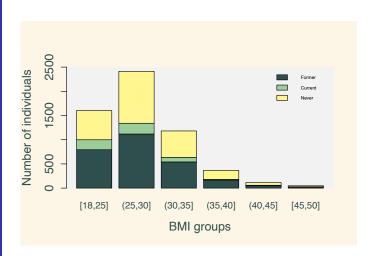


Table of

Outline

Histograms

KDEs

Scatterplo

Royale

Bar plots

Mosaic plots

QQ-plots

Heatmap

Boxplot of proportions

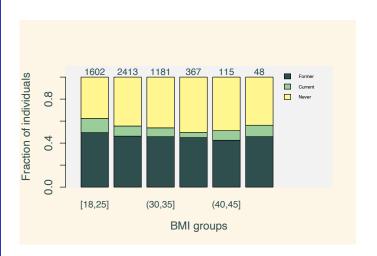


Table of

Outline

Histograms

KDEs

Scatterpio

Болріо

Bar plots

Mosaic plots

QQ-plots

Heatmar

Side-by-side boxplots

Add beside=TRUE to the boxplot function

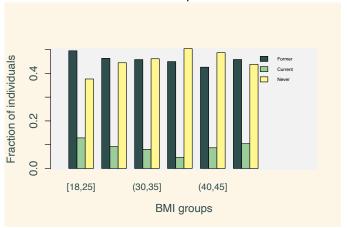


Table of contents

Outillic

Histograms

.....

Scatterplots

Boxplot

Bar plot

Mosaic plots

QQ-plots

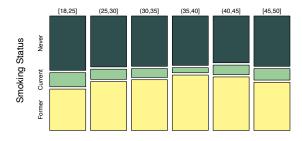
Heatmap:

Mosaic plots

- Mosaic plots are useful for displaying contingency table data
- They are identical to stacked bar plots of proportions

Mosaic plots





BMI groups

Scatterplot

Boxplot

Bar plots

Mosaic plots

QQ-plots

Heatmaps

- QQ-plots (for quantile-quantile) are extremely useful for comparing data to a theoretical distribution
- Plot the empirical quantiles against theoretical quantiles
- Most useful for diagnosing normality

Table of contents

...

listograms

KDE

Scatterplot

2

_ .

Manadamilan

QQ-plots

Heatman

• Let x_p be the p^{th} quantile from a $N(\mu, \sigma^2)$

- Then $P(X \le x_p) = p$
- Clearly $P(Z \le \frac{x_p \mu}{\sigma}) = p$
- Therefore $x_p = \mu + z_p \sigma$ (this should not be news)
- Result: quantiles from a $N(\mu, \sigma^2)$ population should be linearly related to standard normal quantiles
- A normal qq-plot plot displays the empirical quantiles against the theoretical standard normal quantiles
- In R qqnorm for a normal QQ-plot and qqplot for a qqplot against an arbitrary distribution

Table of

Outille

Histogram:

KDE:

Scatterplo

Boynla

Bar plot

Mosaic plots

QQ-plots

QQ-plot for BMI versus Normal

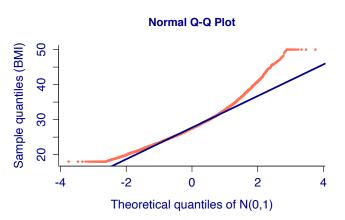


Table of

0 4 4 1 1 1 1

Histograms

KDE:

Scatterplo

Dotcharts

Boxplo

Bar plots

iviosaic pioi

QQ-plots

and the second

QQ-plot for age versus Normal

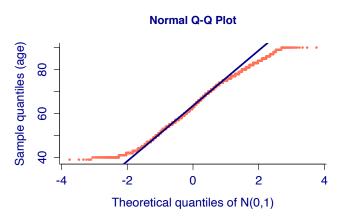


Table of contents

Outille

Histograms

NDES

Scatterplo

Dotcharts

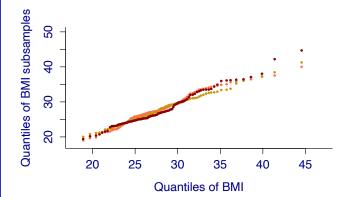
Roxbio

Dai piots

QQ-plots

11.

QQ-plots for three subsamples of size 100 versus BMI quantiles



miscogram:

INDES

Scatterplot

Boxplot

Bar plot

Mosaic plots

QQ-plots

Heatmaps

- Display of the data in matrix format
- Each axis: a variable or study participants by variables
- Intensity of the color: values taken at that particular pair of variables
- Example: temperature at a particular latitude/longitude
- Example: BMI for a study participant

Table of

Outillic

Histogram:

KDE

Scatterplot

Doxplots

.

mosare pro-

QQ-plot

Heatmaps

Heatmap of correlation

```
library(corrplot)
subset.data.cv=data.cv[,c(2,4,10:11,24,26:27,29)]
M=cor(subset.data.cv,use="pairwise.complete.obs")
corrplot(M, method = "square")
```

Table of

Outline

Histograms

KDF

Scatterplot

Roxblot

--- ----

iviosaic pio

QQ-plots

Heatmaps

Correlations among eight variables in SHHS

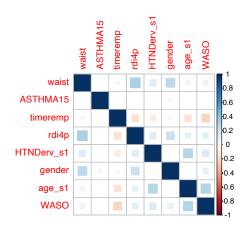


Table of contents

Outille

Histogram

KDES

Scatterpioi

2

_ .

00 1.

Heatmaps

Correlations using the fields package

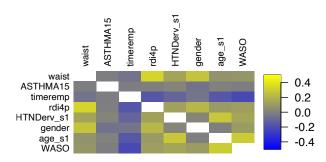


Table of

Outline

Histograms

KDEs

Scatterplot

Dotchart

Mosaic plot

QQ-plots

Heatmaps

Heatmap all observations, standardized

