Lecture 11

Ciprian M. Crainiceanu

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

August 21, 2021

Table of contents

Outline

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

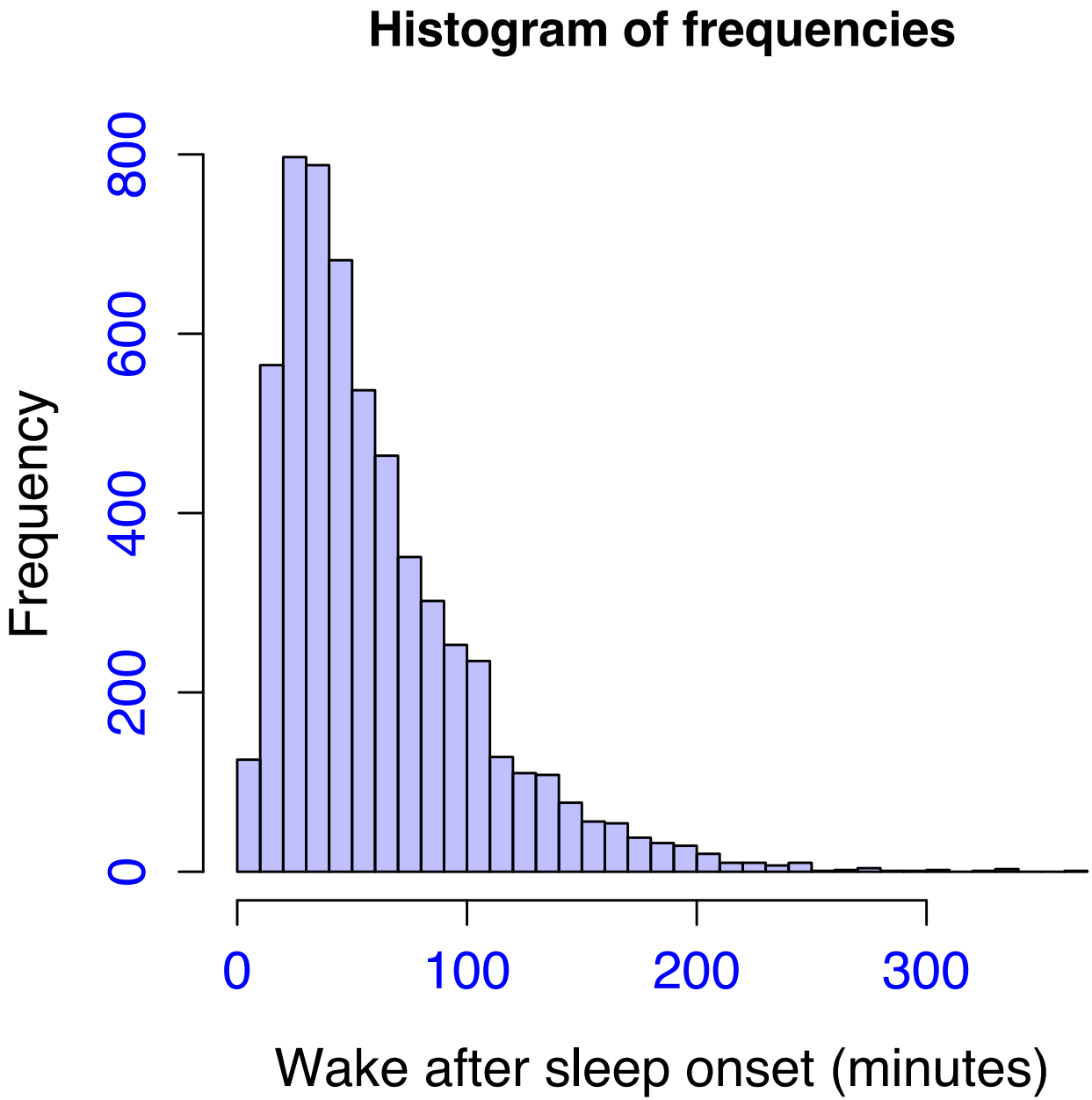
Histograms

- 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

Example: Frequency histograms

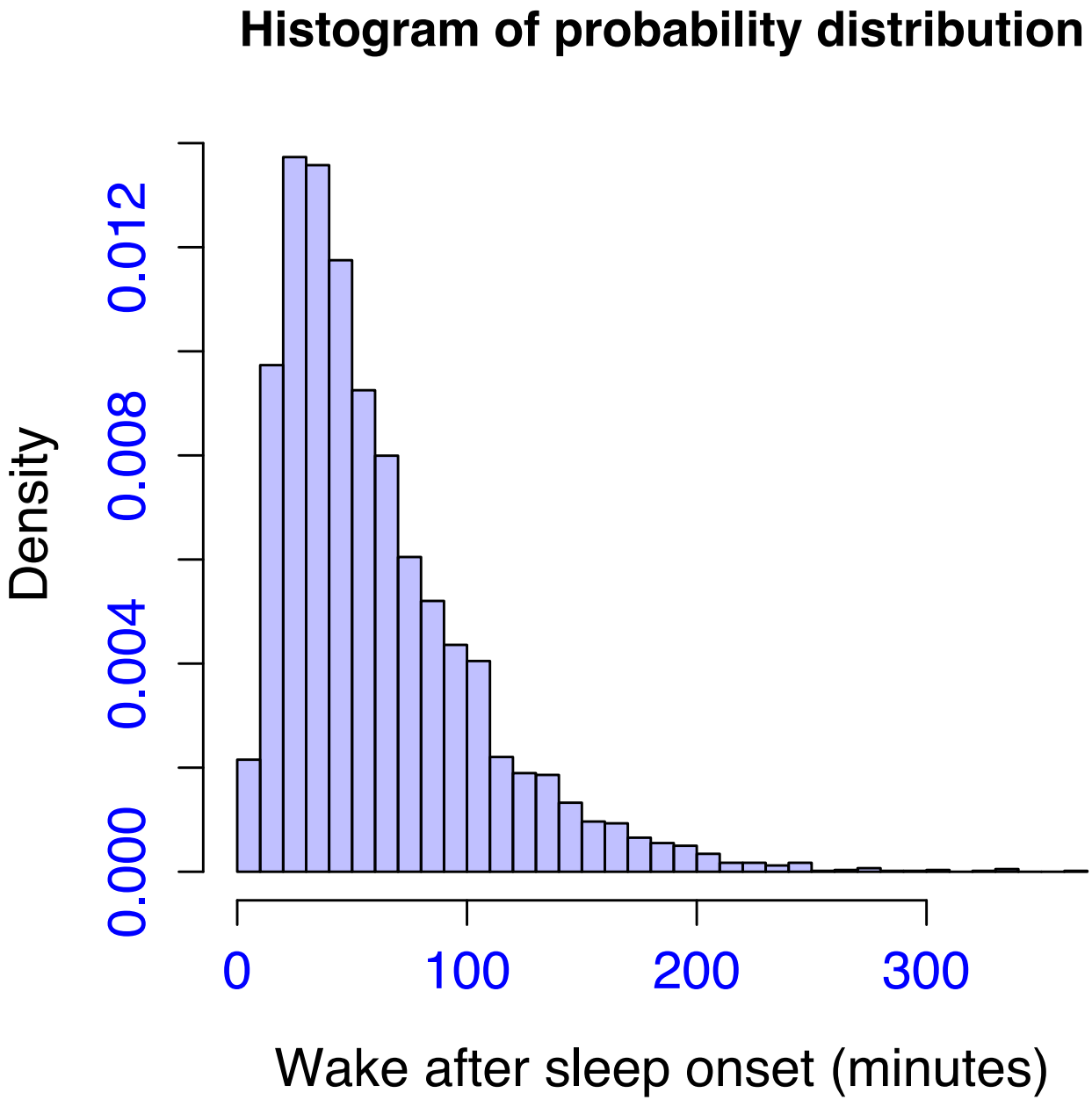
- Sleep Heart Health Data
- WASO (Wake After Sleep Onset): linked to sleep quality.
- Expressed in minutes

```
hist(WASO,col=rgb(0,0,1,1/4),breaks=30,
     xlab="Wake after sleep onset (minutes)",
     main="Histogram of frequencies",
     cex.lab=1.3,cex.axis=1.3,col.axis="blue")
```



Example: Probability distribution histograms

```
hist(WASO,probability=TRUE,col=rgb(0,0,1,1/4),
     breaks=30,
     xlab="Wake after sleep onset (minutes)",
     main="Histogram of probability distribution",
     cex.lab=1.3,cex.axis=1.3,col.axis="blue")
```



Pros and cons

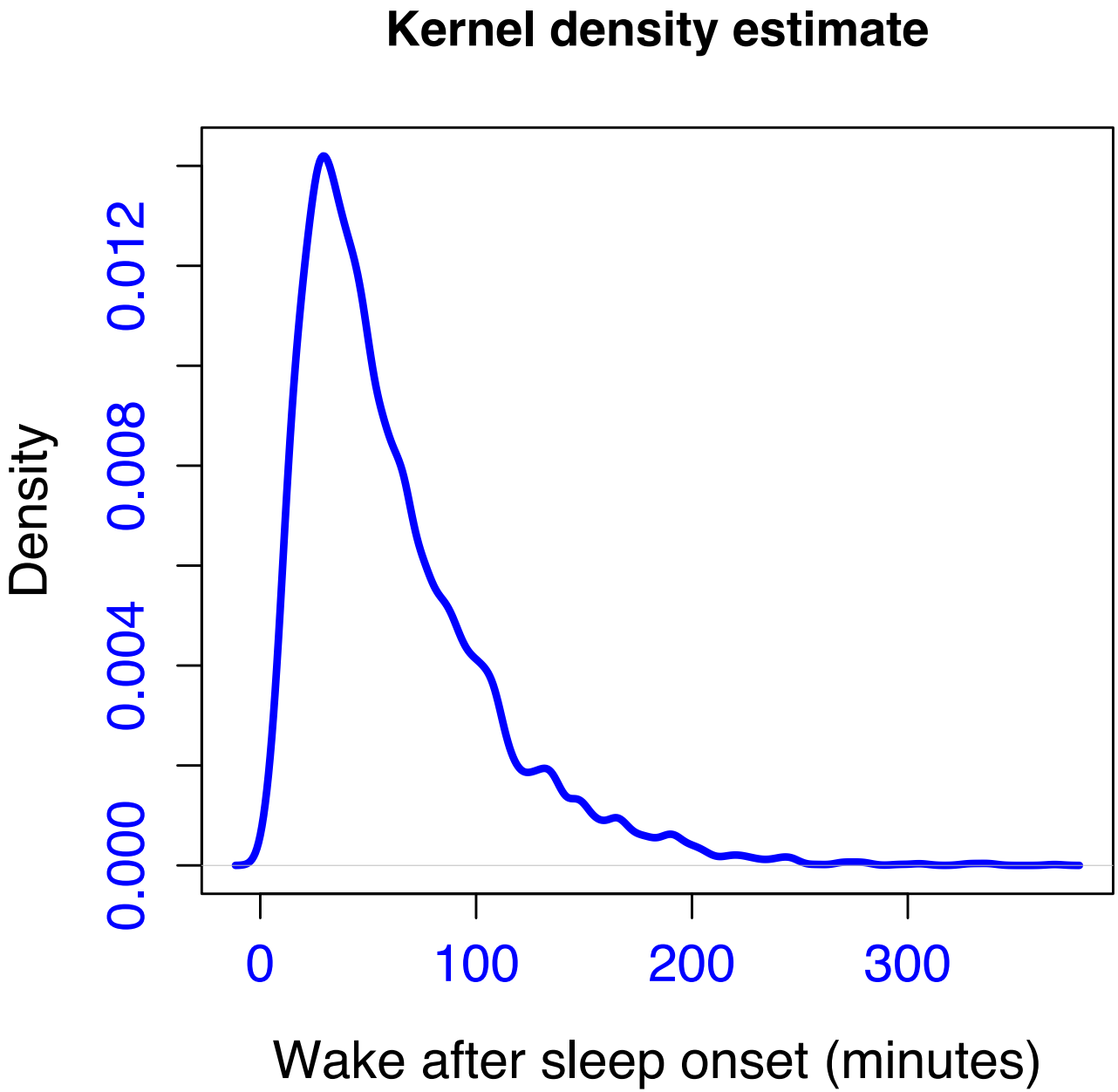
- Histograms are useful and easy, apply to continuous, discrete and even unordered data
- They use a lot of ink and space to display information
- It is difficult to display several at the same time
- Certain distributions may require data transformation for proper plotting

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
  a. 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

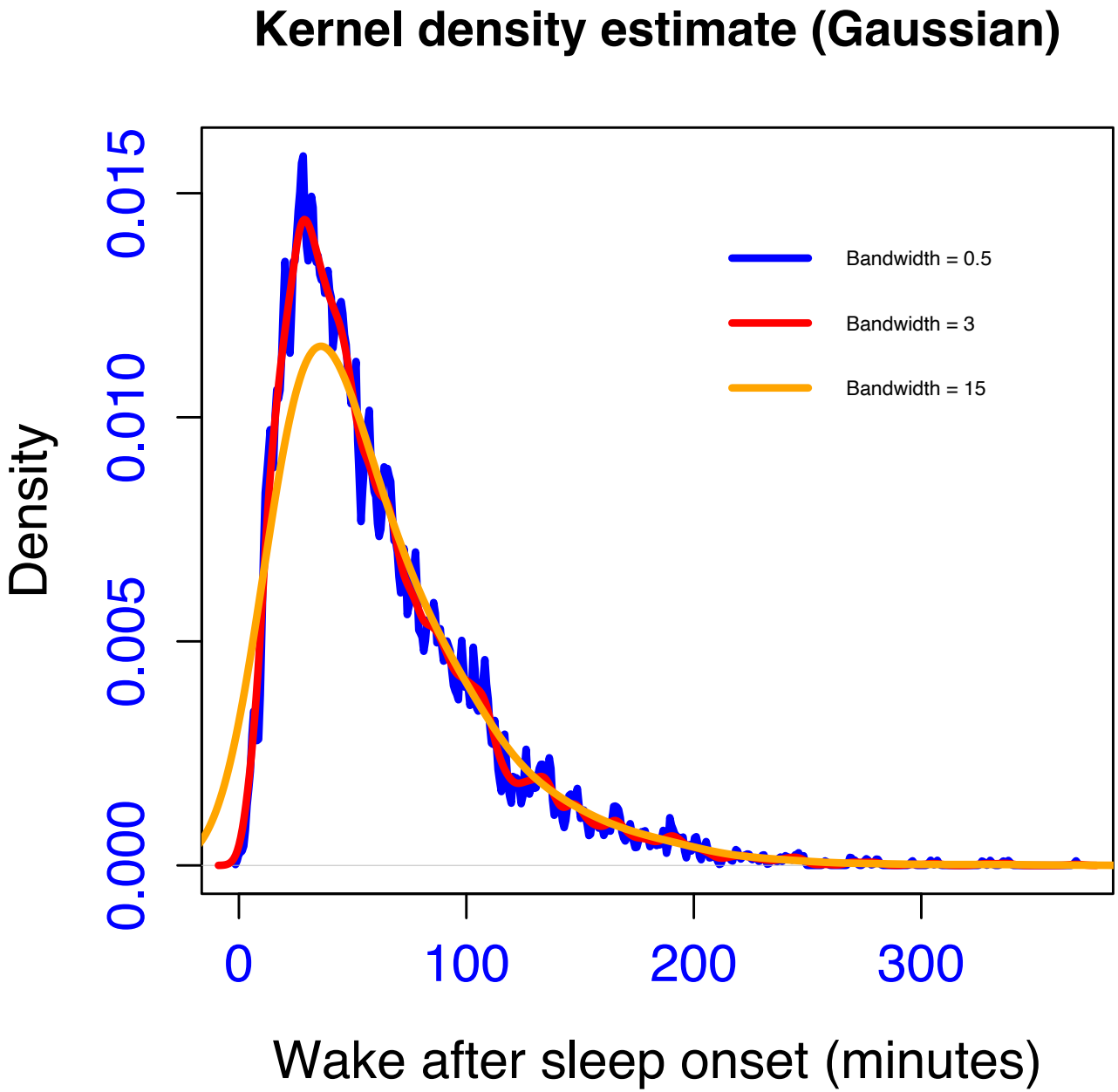
Example: Automatic KS

```
d<-density(WASO,bw="sj")
plot(d,col="blue",lwd=3,
     xlab="Wake after sleep onset (minutes)",
     main="Kernel density estimate",
     cex.lab=1.3,cex.axis=1.3,col.axis="blue")
```



Example: Automatic KS

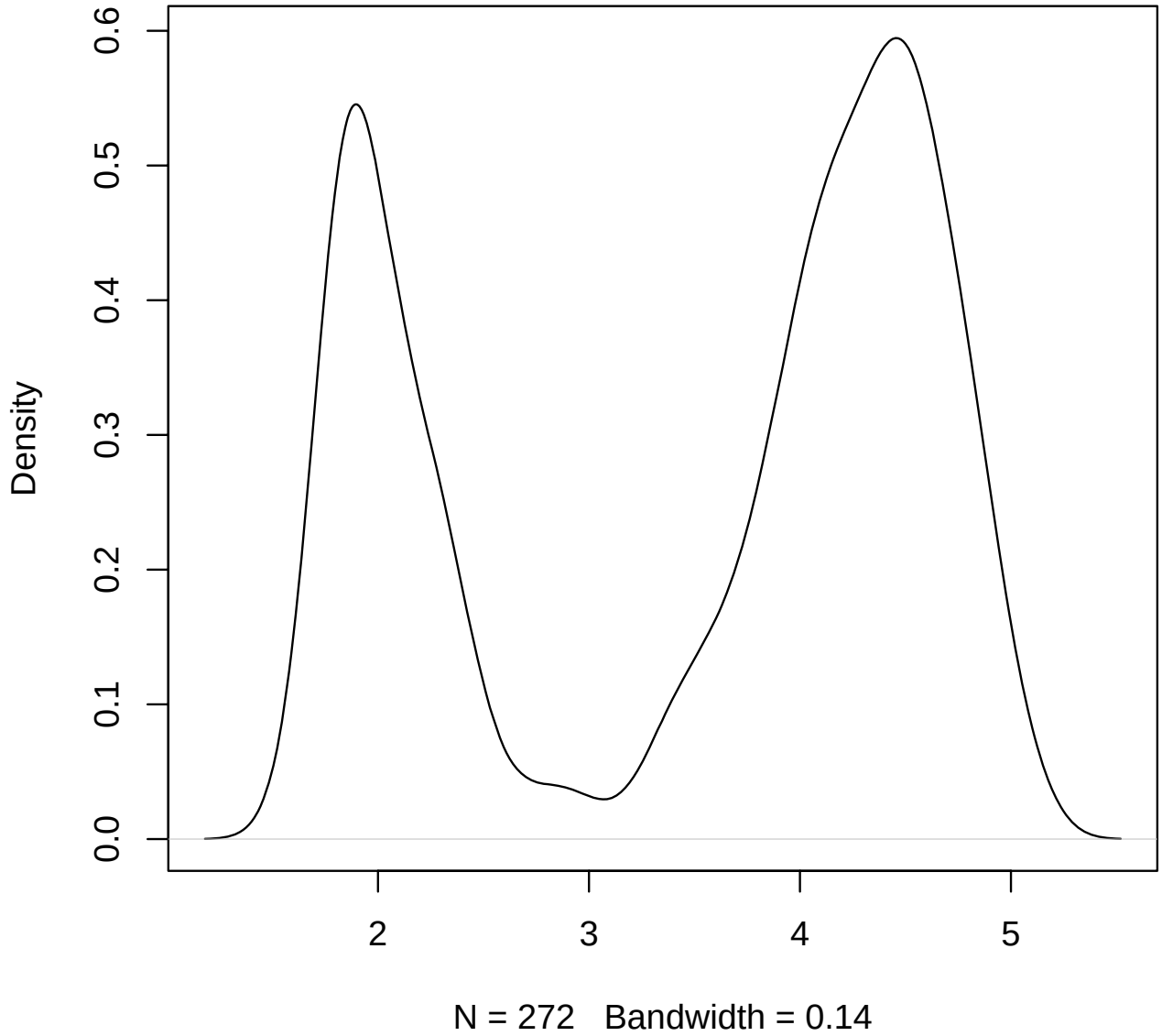
```
d1<-density(WASO,bw=0.5)
d2<-density(WASO,bw=3)
d3<-density(WASO,bw=15)
plot(d1,col="blue",lwd=3,
     xlab="Wake after sleep onset (minutes)",
     main="Kernel density estimate (Gaussian)",
     cex.lab=1.3,cex.axis=1.3,col.axis="blue")
lines(d2,lwd=3,col="red")
lines(d3,lwd=3,col="orange")
```



Example

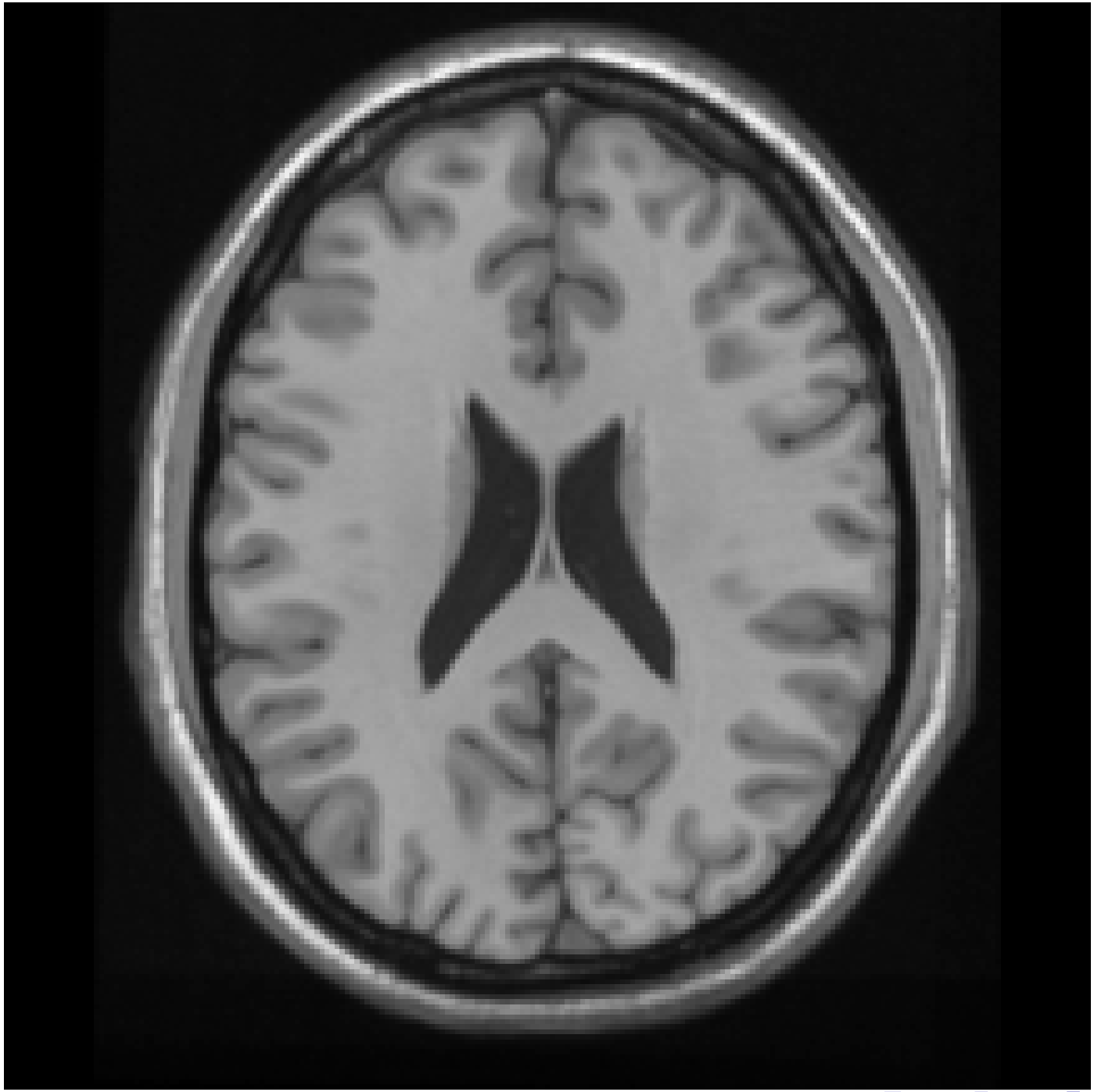
Data is the waiting and eruption times in minutes between eruptions of the Old Faithful Geyser in Yellowstone National park

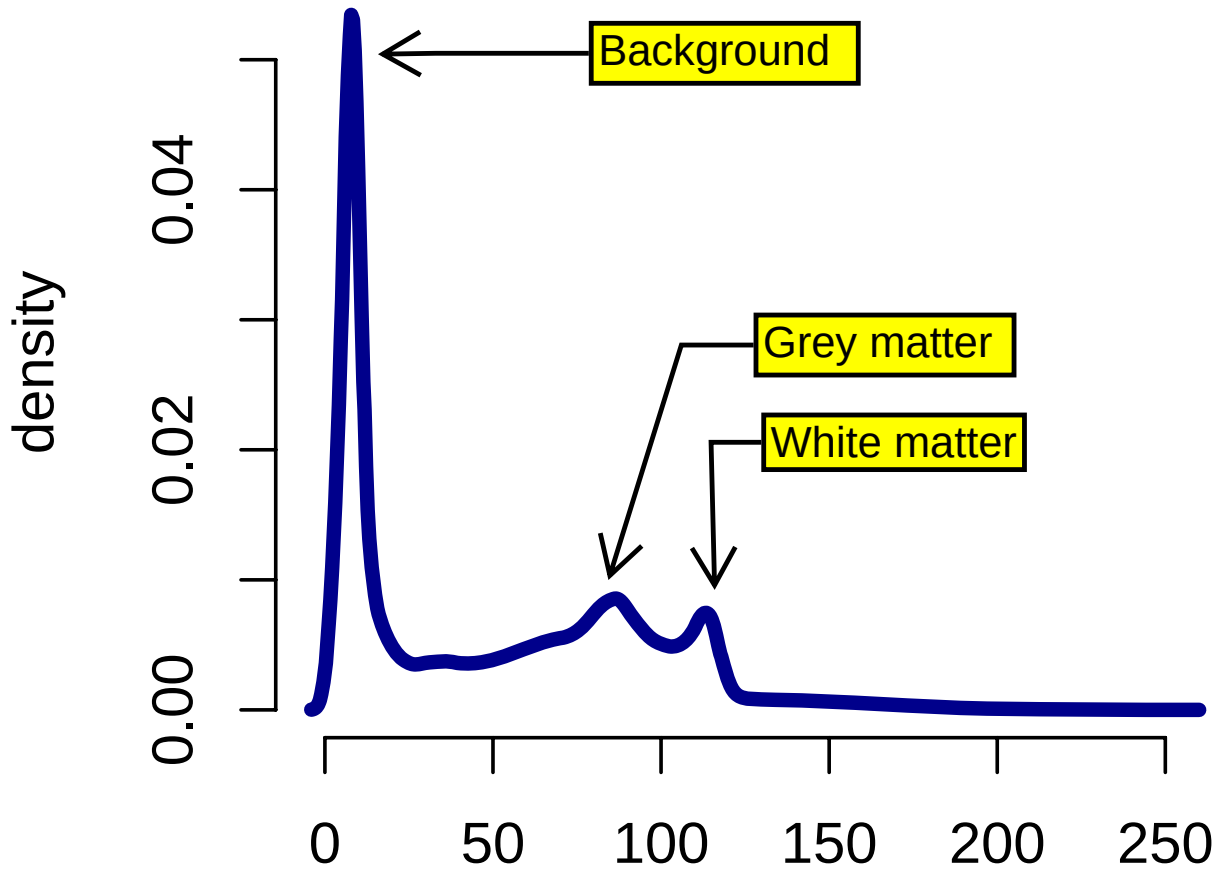
```
data(faithful)
d <- density(faithful$eruptions, bw = "sj")
plot(d)
```



Imaging example

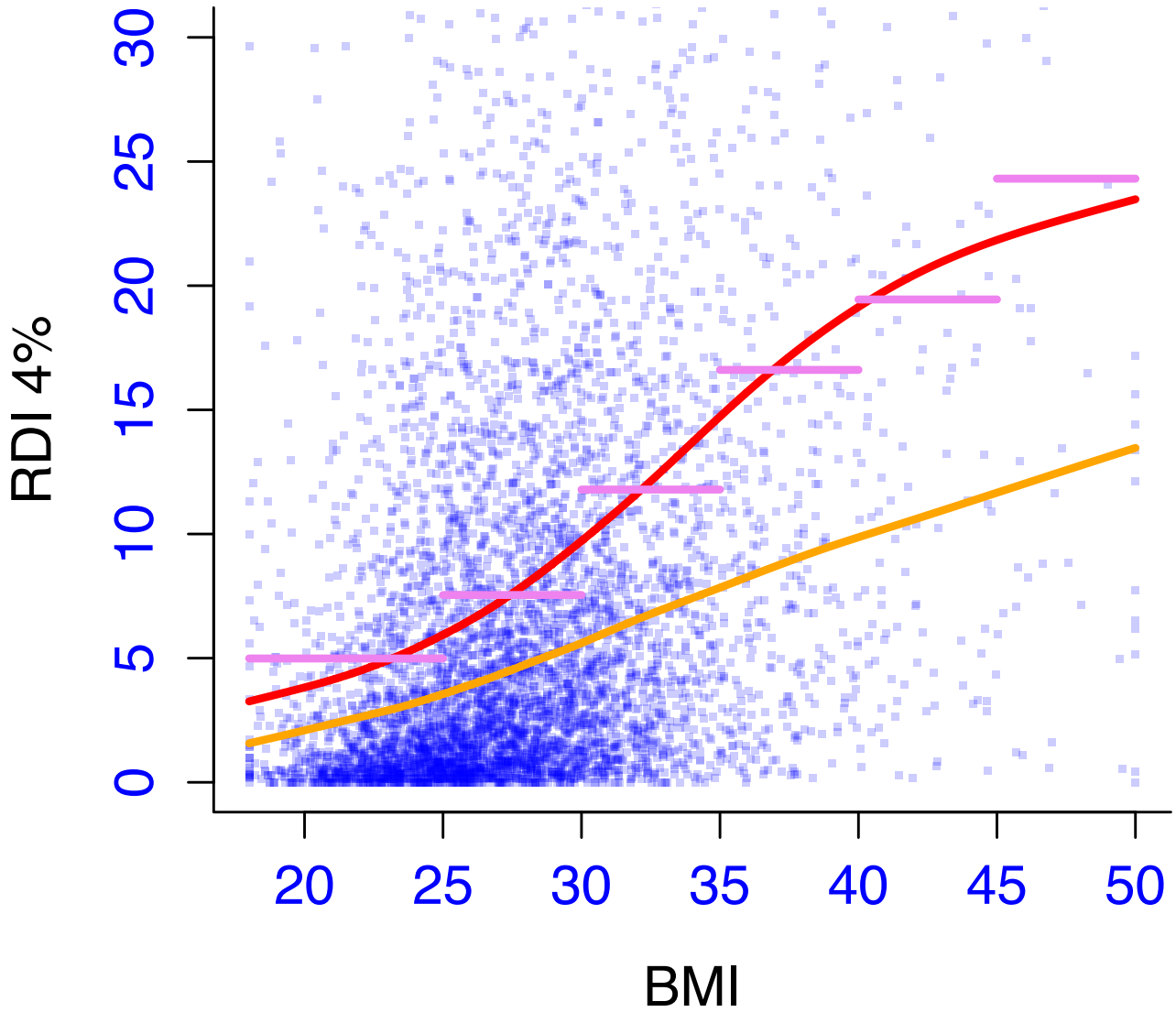
- Consider the following image slice (created in R) from a high resolution MRI of a brain
- This is a single (axial) slice of a three-dimensional image
- Consider discarding the location information and plotting a KDE of the intensities





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

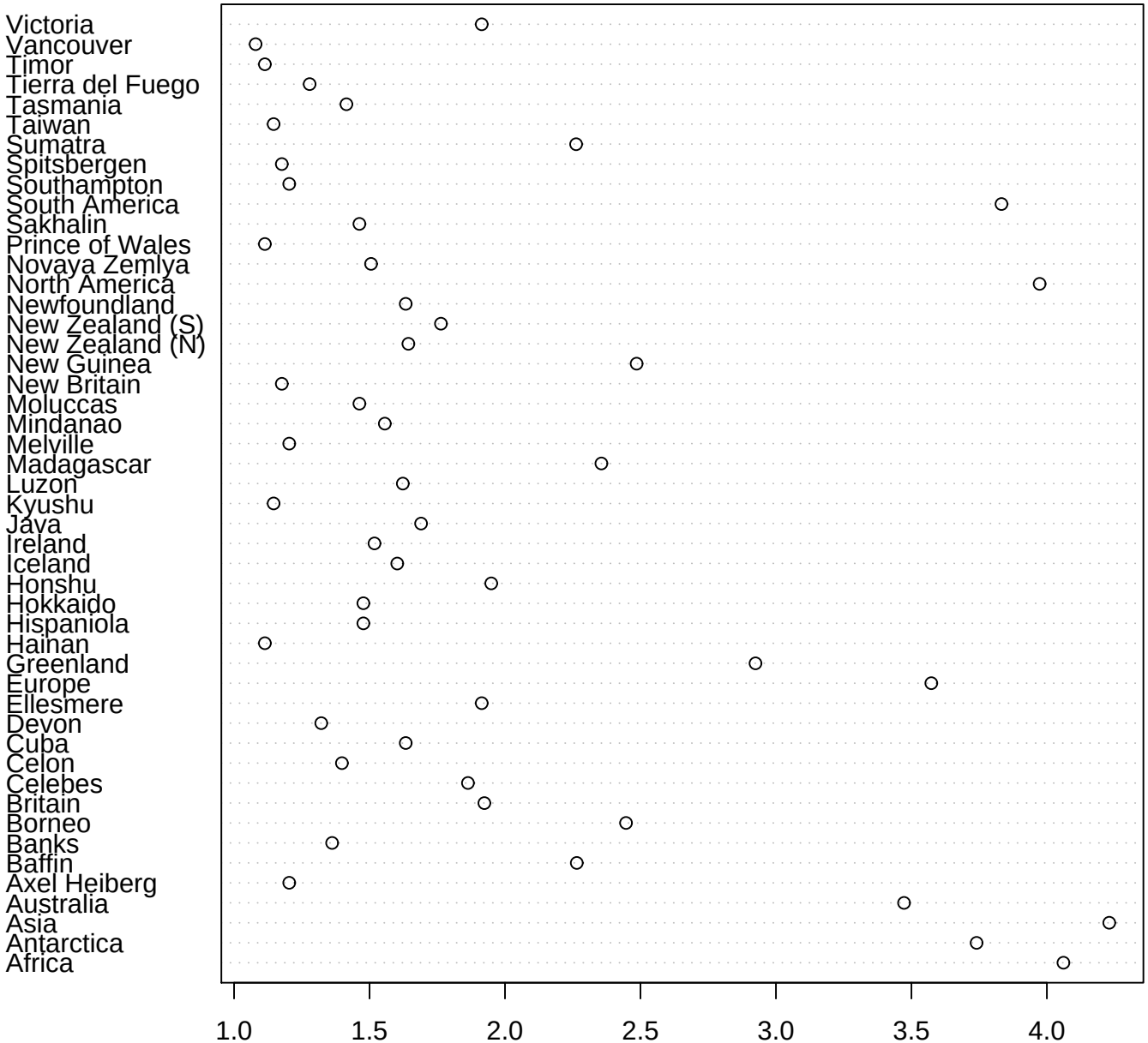


Dotcharts

- 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))
```

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



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)?

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`)

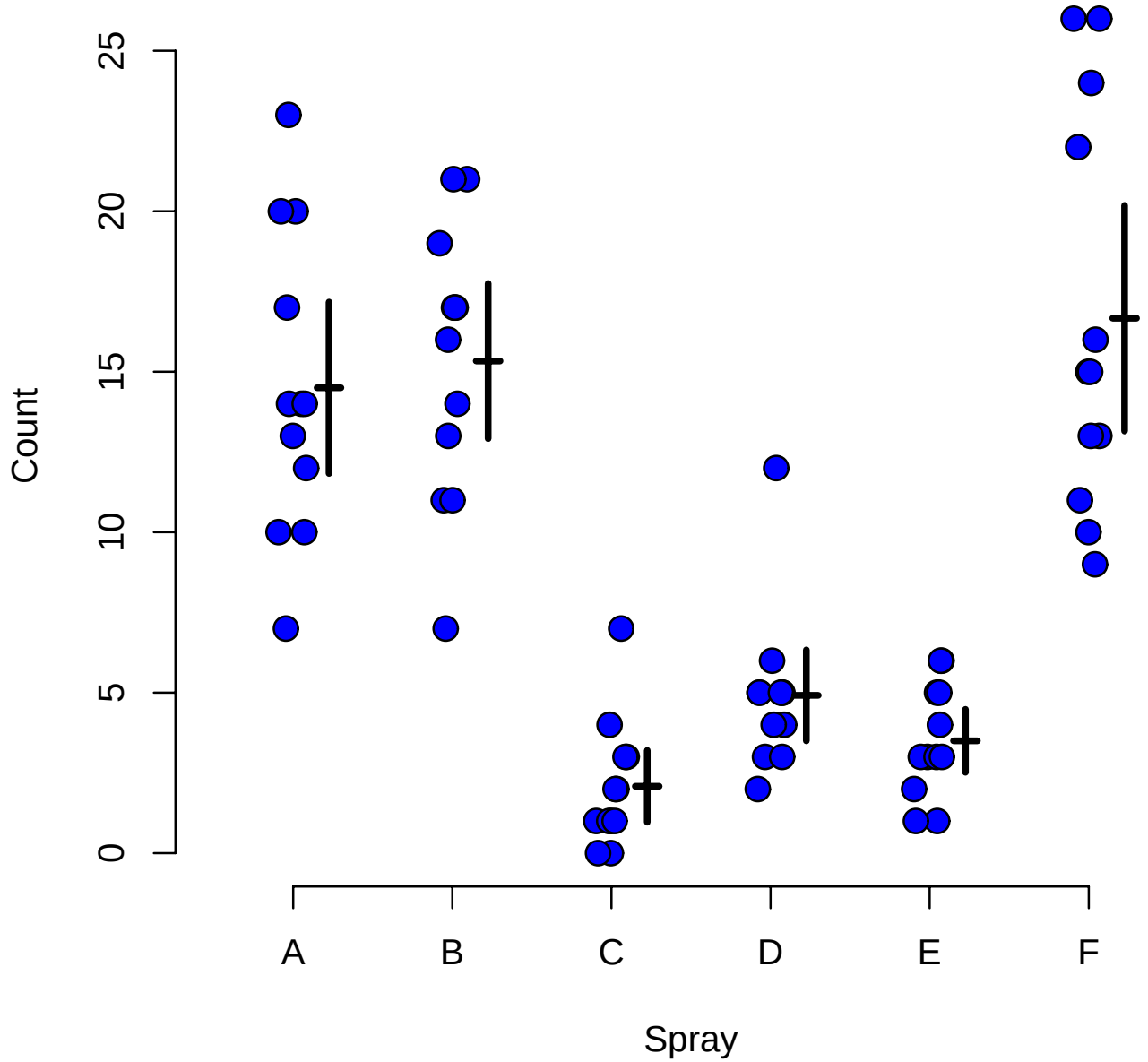
Example

- 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)
```

The gist of the code is below

```
attach(InsectSprays)
plot(c(.5, 6.5), range(count))
sprayTypes <- unique(spray)
for (i in 1 : length(sprayTypes)){
  y <- count[spray == sprayTypes[i]]
  n <- sum(spray == sprayTypes[i])
  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)
       )
}
```



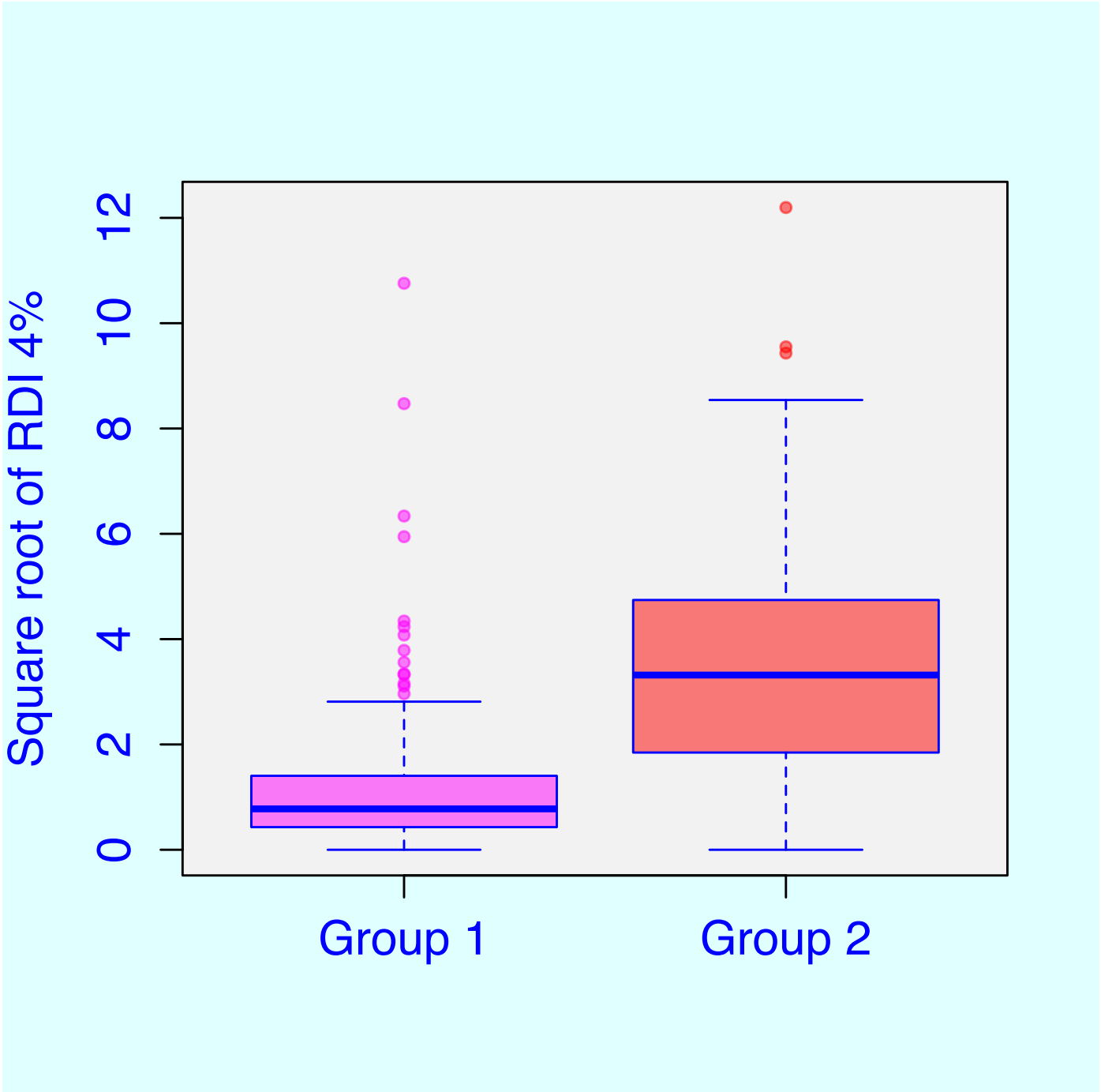
Boxplots

- 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

Example: SHHS

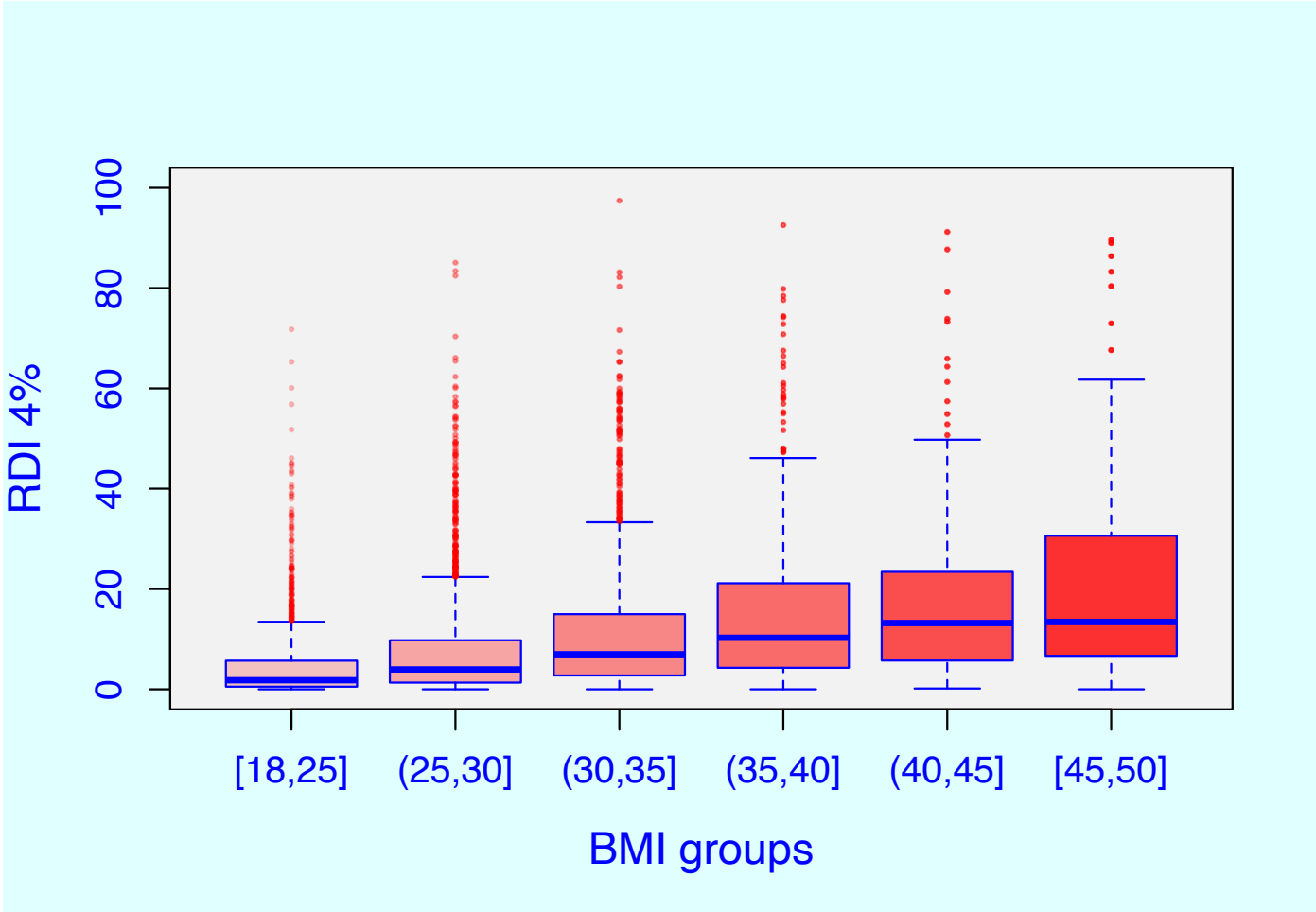
- Distribution of RDI 4% in two groups.
- Group 1: BMI$<$25 and age$<$50
- Group 2: BMI$>$35 and age$>$70

```
data.box<-list(sqrt_rdi_g1,sqrt_rdi_g2)
boxplot(data.box,col=c(col=rgb(1,0,1,0.5),
        col = rgb(1,0,0,0.5)),cex.lab=1.3,
        cex.axis=1.3,col.axis="blue",col.lab="blue",
        names=c("Group 1","Group 2"),
        ylab="Square root of RDI 4%",
        border=c("blue","blue"),
        outpch=20,outcex=1,
        outcol=c(col=rgb(1,0,1,0.5),
        col = rgb(1,0,0,0.5)))
```



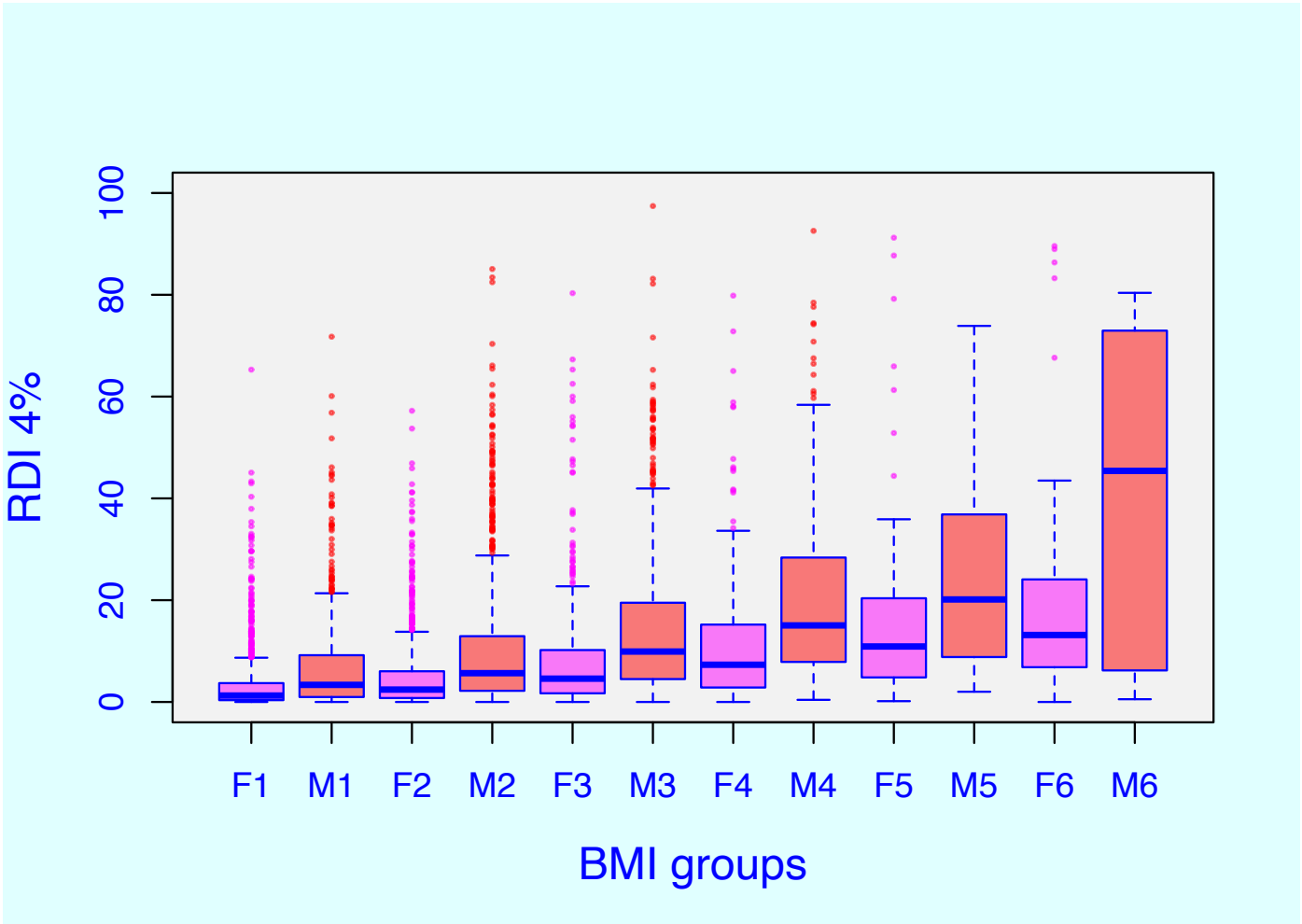
Example: SHHS

- Distribution of RDI 4% in six BMI groups.



Example: SHHS

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



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

Bar plots

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

Data structure for a simple bar plot

```
counts <- table(bmi_cut)
```

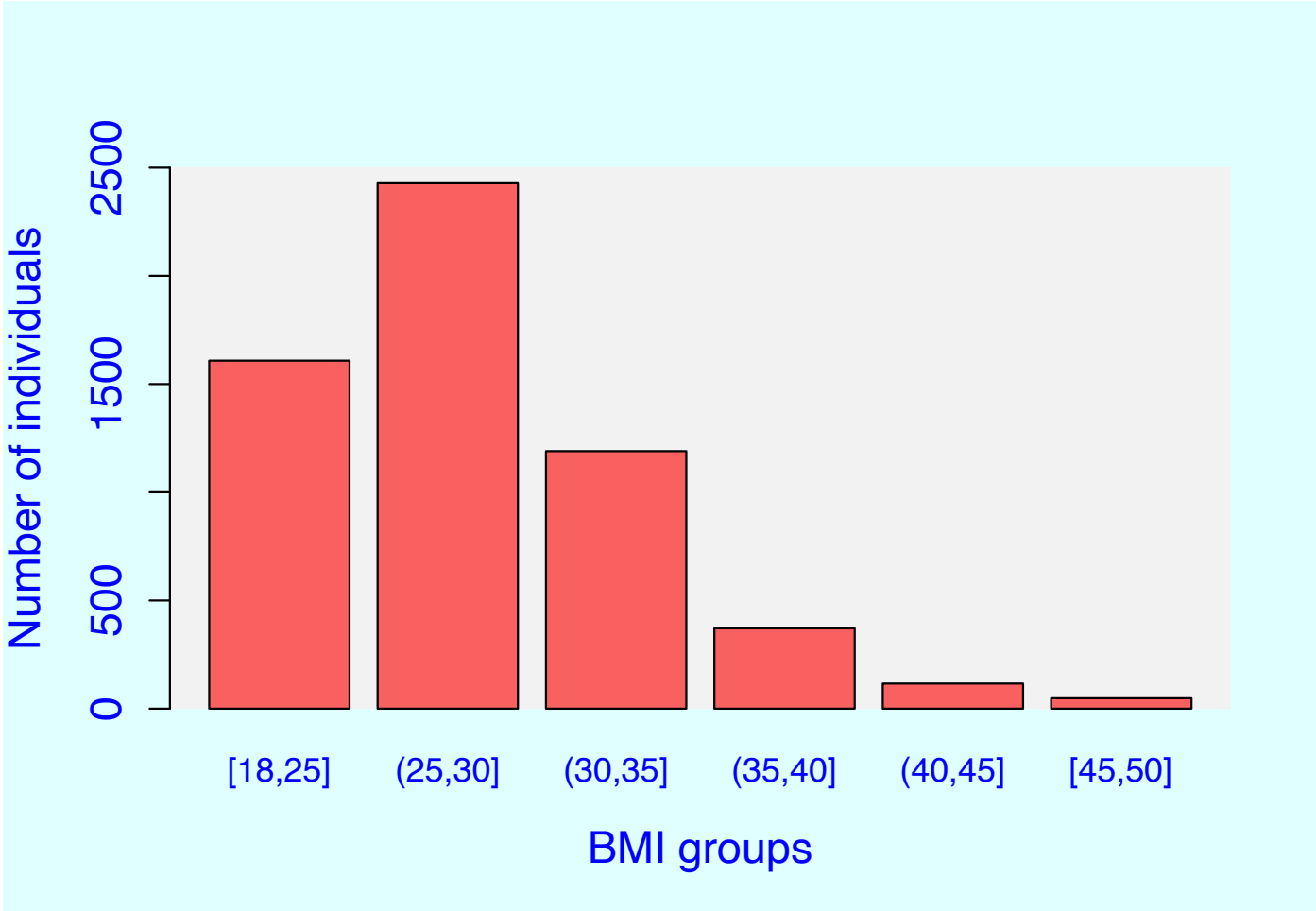
```
> counts
          bmi_cut
[18,25] (25,30] (30,35] (35,40] (40,45] [45,50]
   1608    2428    1190     371     116      48
```

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")
rect(u[1],u[3],u[2],u[4], col="gray95", border=NA)
par(new=TRUE)

barplot(counts, main="",
    ylab="Number of individuals",
    xlab="BMI groups",col=rgb(1,0,0,0.6),
    ylim=c(0,2500),cex.axis=1.3,
    col.axis="blue",cex.lab=1.3,
    col.lab="blue")
```

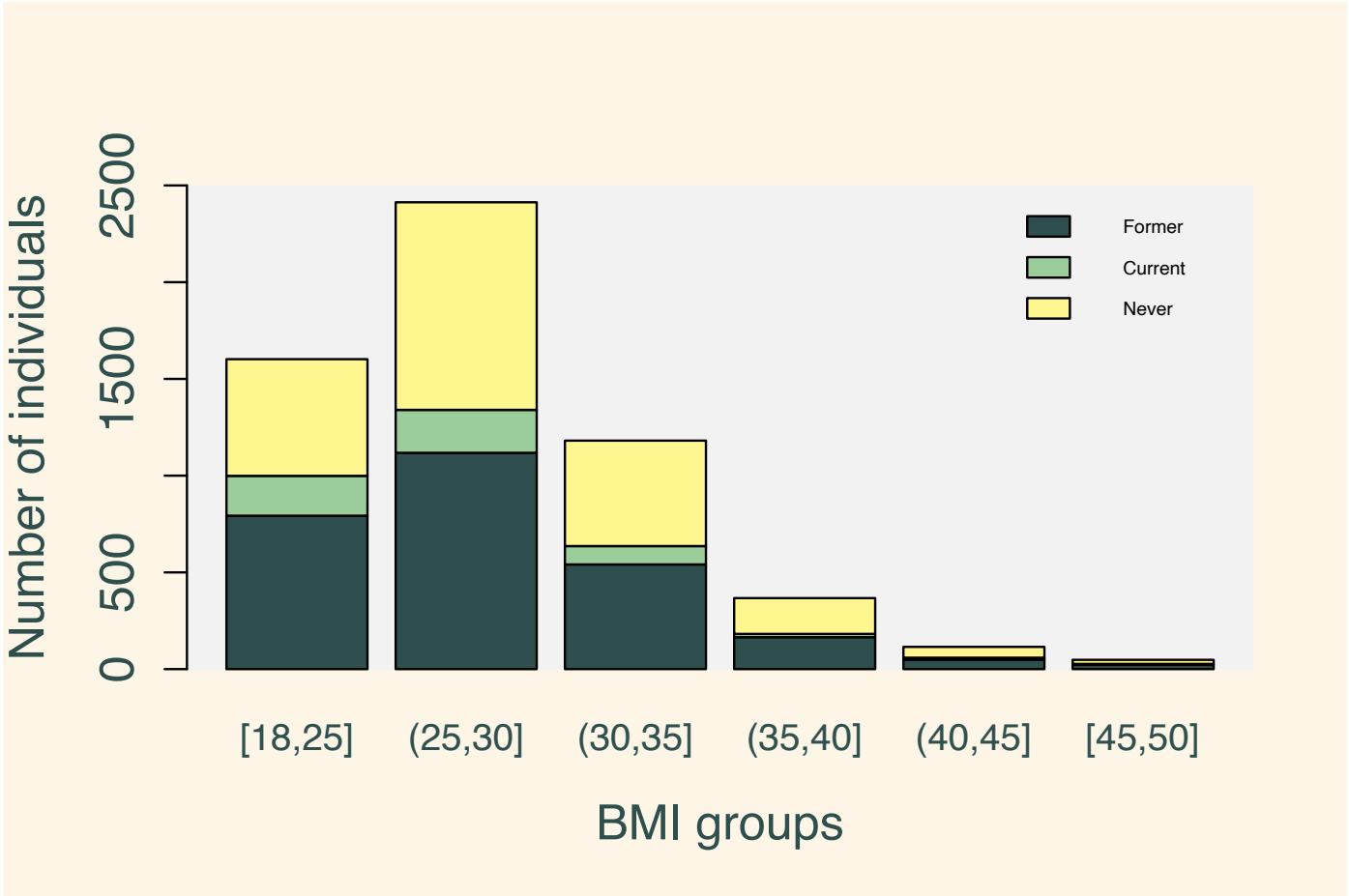


Data structure for a stacked bar plot

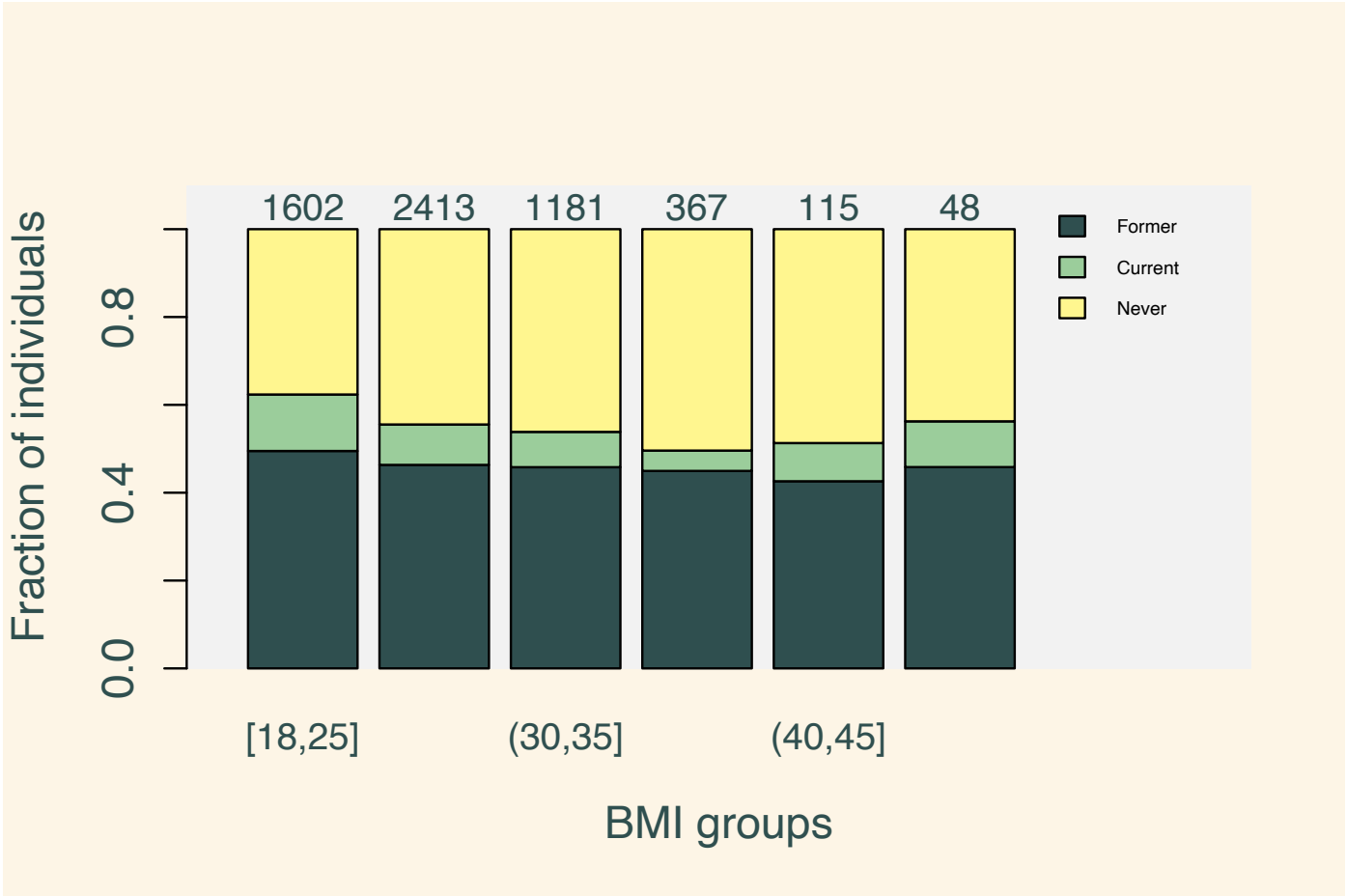
```
counts <- table(smokstatus,bmi_cut)
```

```
> 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
```

Boxplot of counts

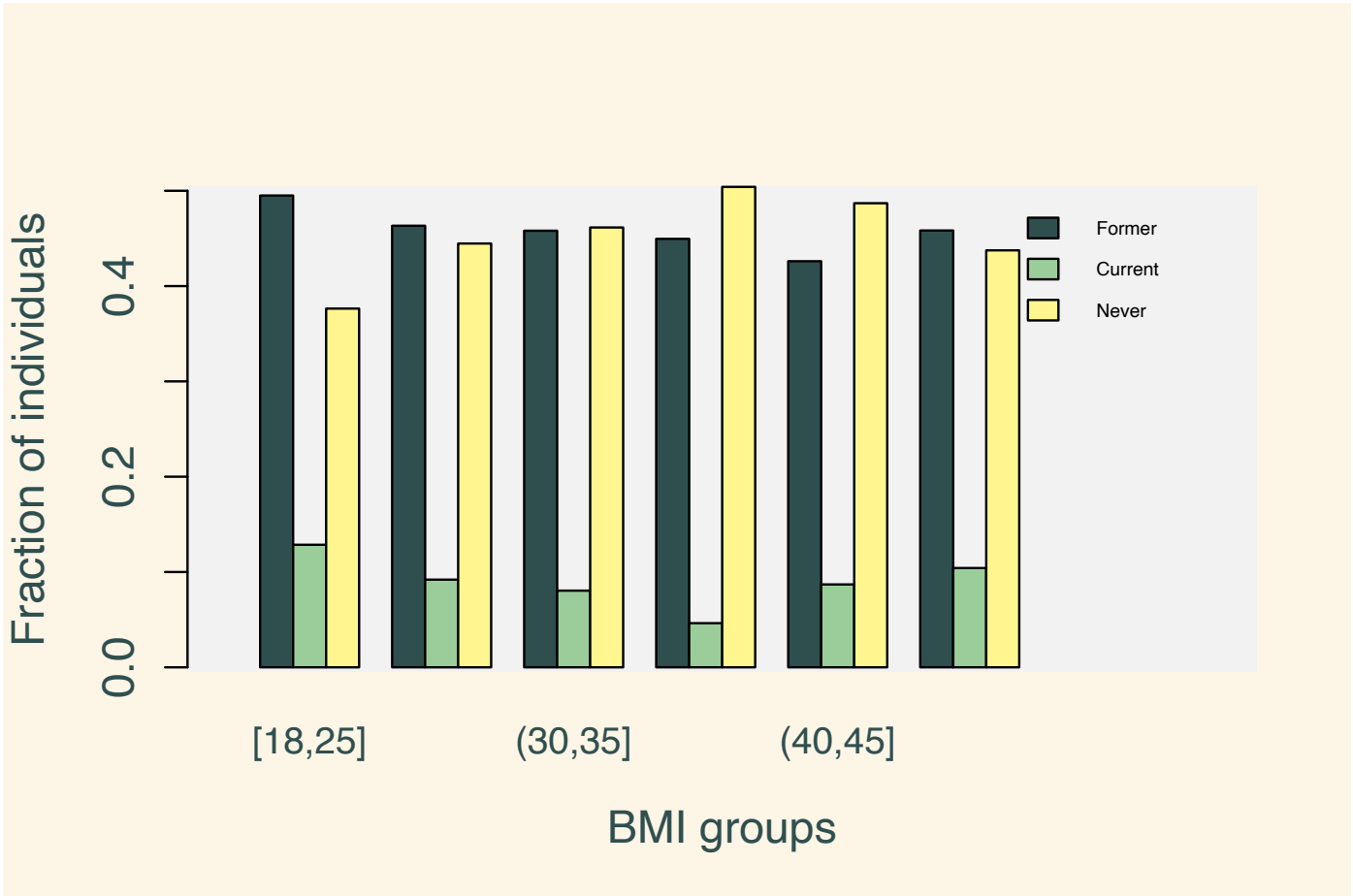


Boxplot of proportions



Side-by-side boxplots

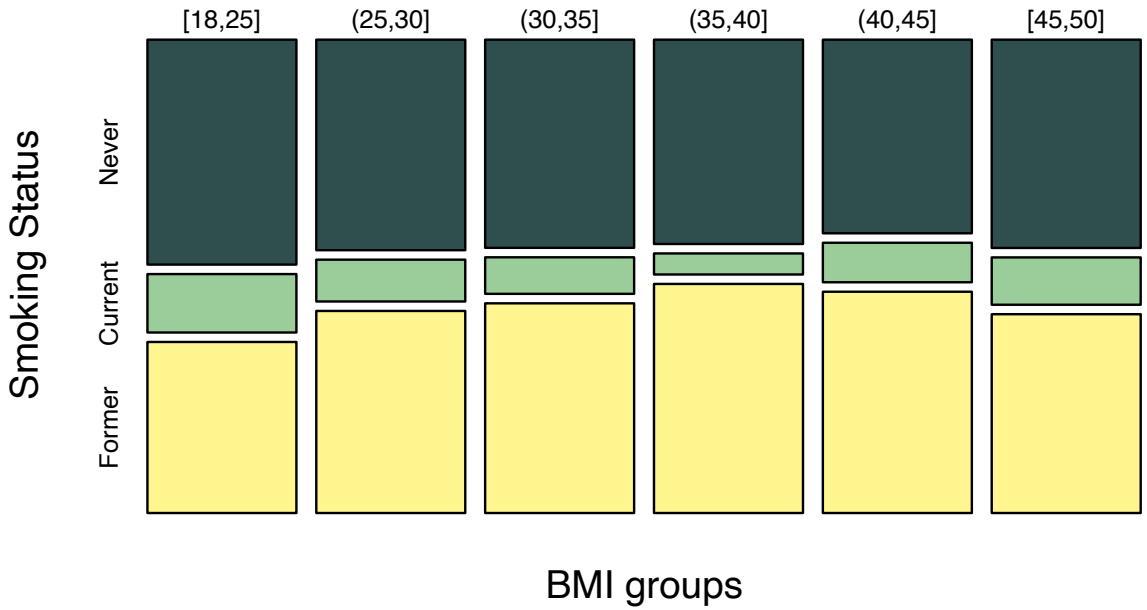
Add `beside=TRUE` to the boxplot function



Mosaic plots

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

```
mosaicplot(t(propt),
           col=c("darkslategray","darkseagreen3","khaki1"),
           xlab="BMI groups",
           ylab="Smoking Status",main="")
```

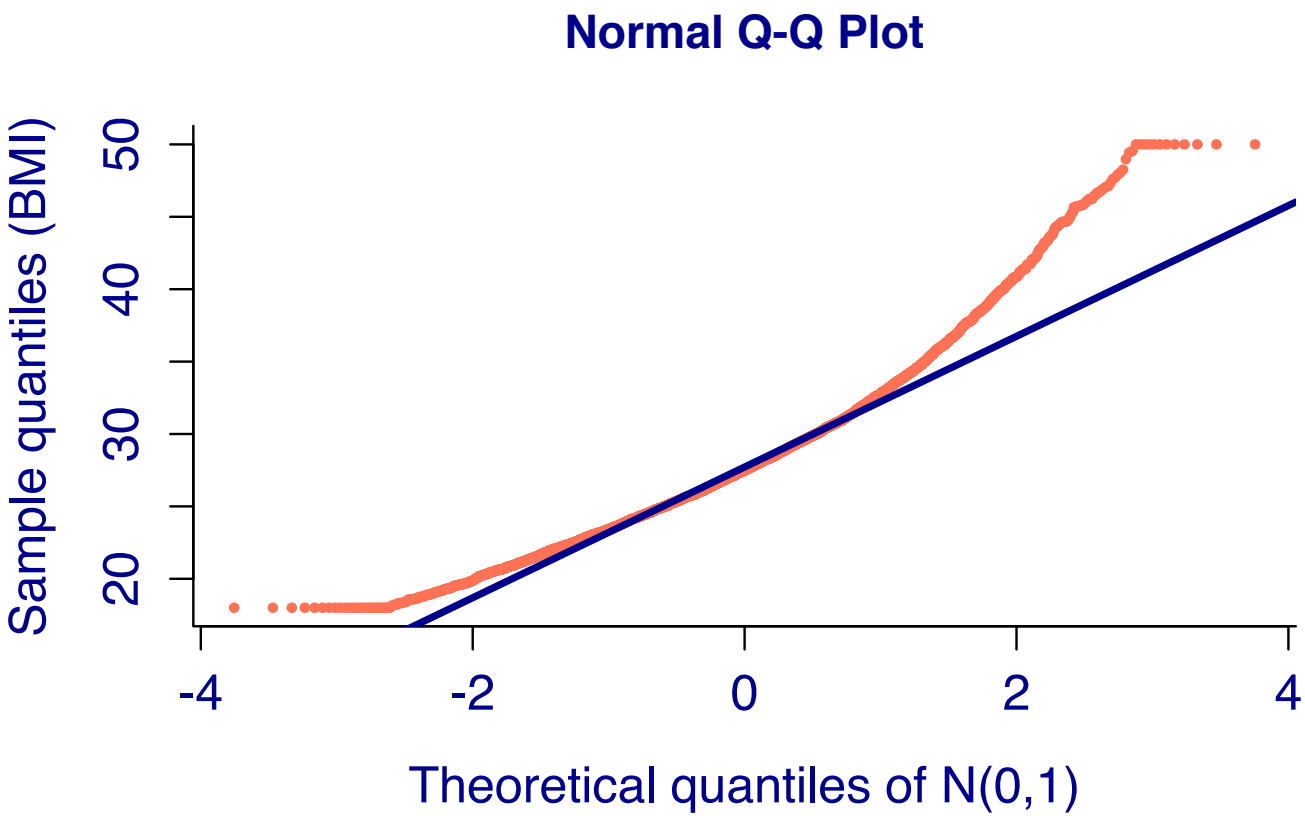


QQ-plots

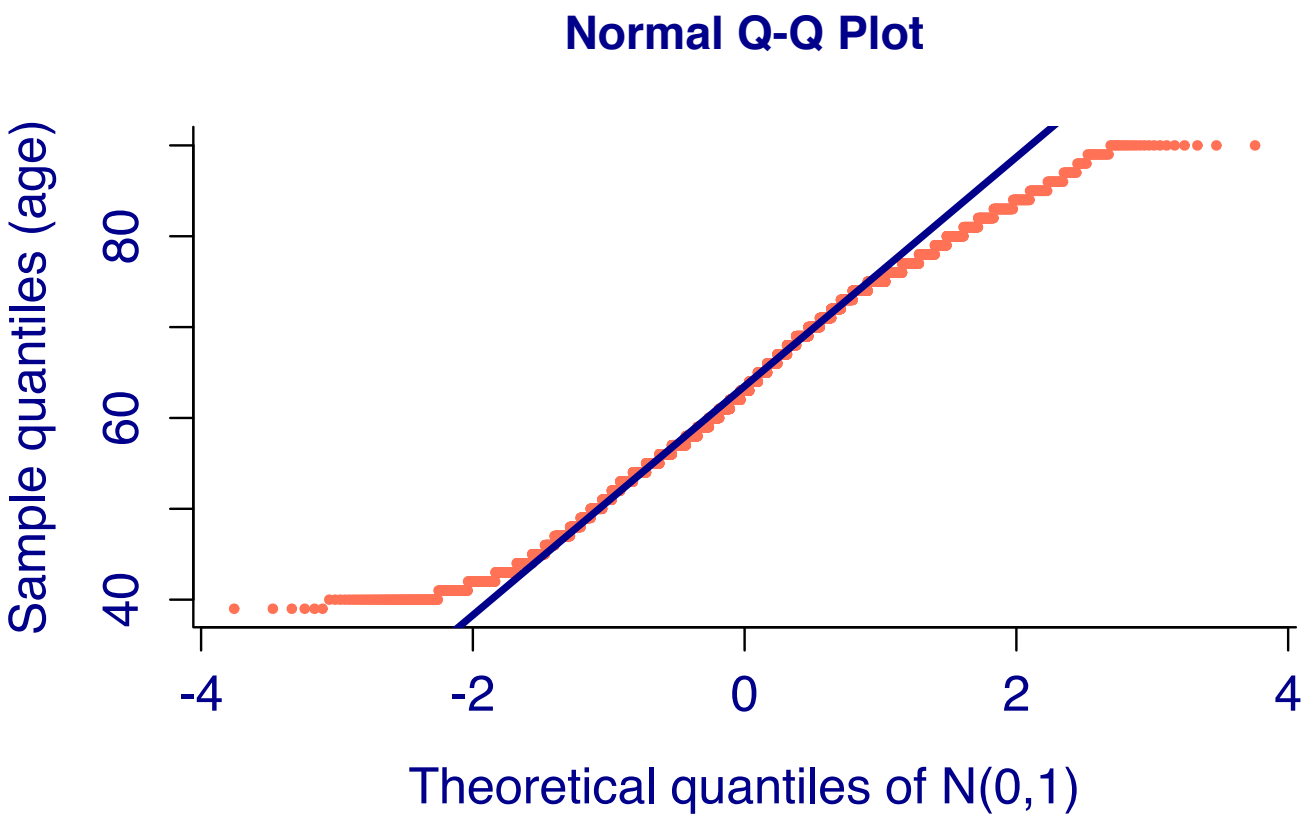
- QQ-plots (for quantile-quantile) are extremely useful for comparing data to a theoretical distribution
- Plot the empirical quantiles against theoretical quantiles
- Most useful for diagnosing normality

- Let x_p be the p^{th} quantile from a $N(\mu, \sigma^2)$
- Then $P(X \leq x_p) = p$
- Clearly $P(Z \leq \frac{x_p - \mu}{\sigma}) = p$
- 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

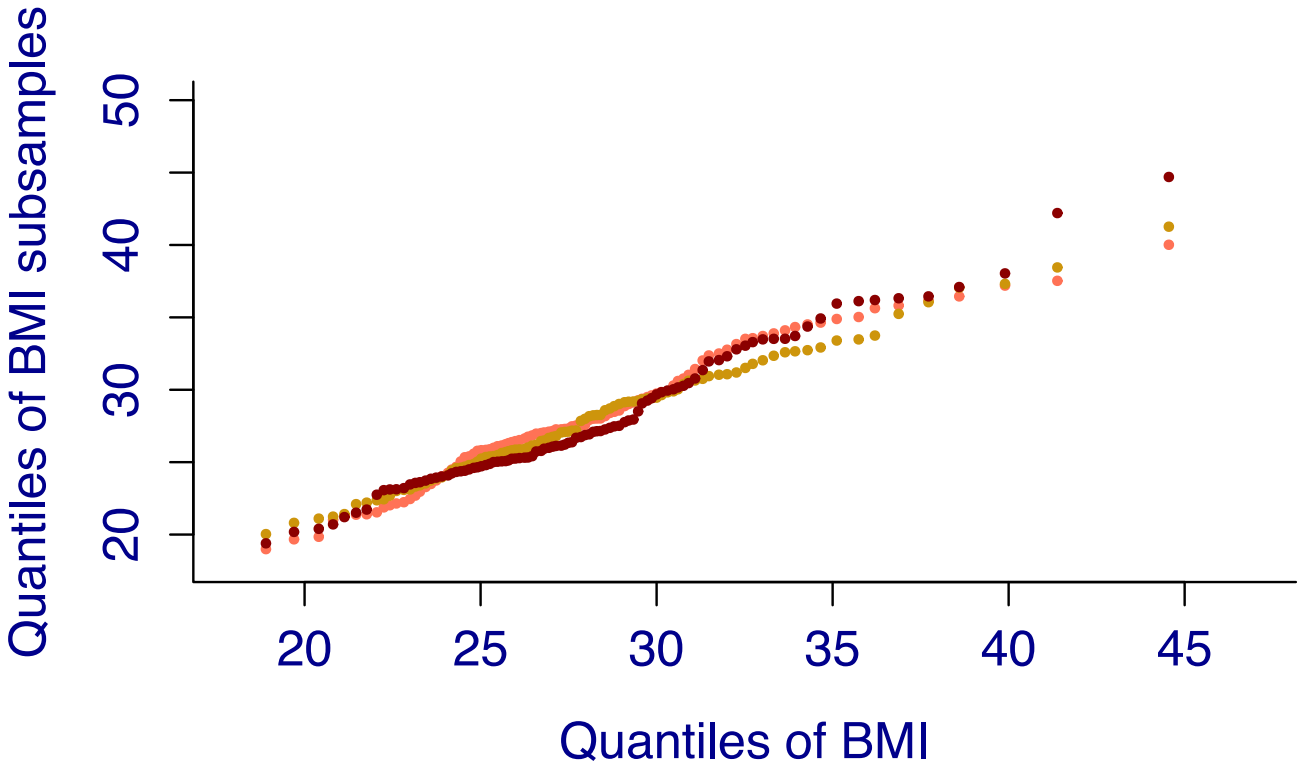
QQ-plot for BMI versus Normal



QQ-plot for age versus Normal



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



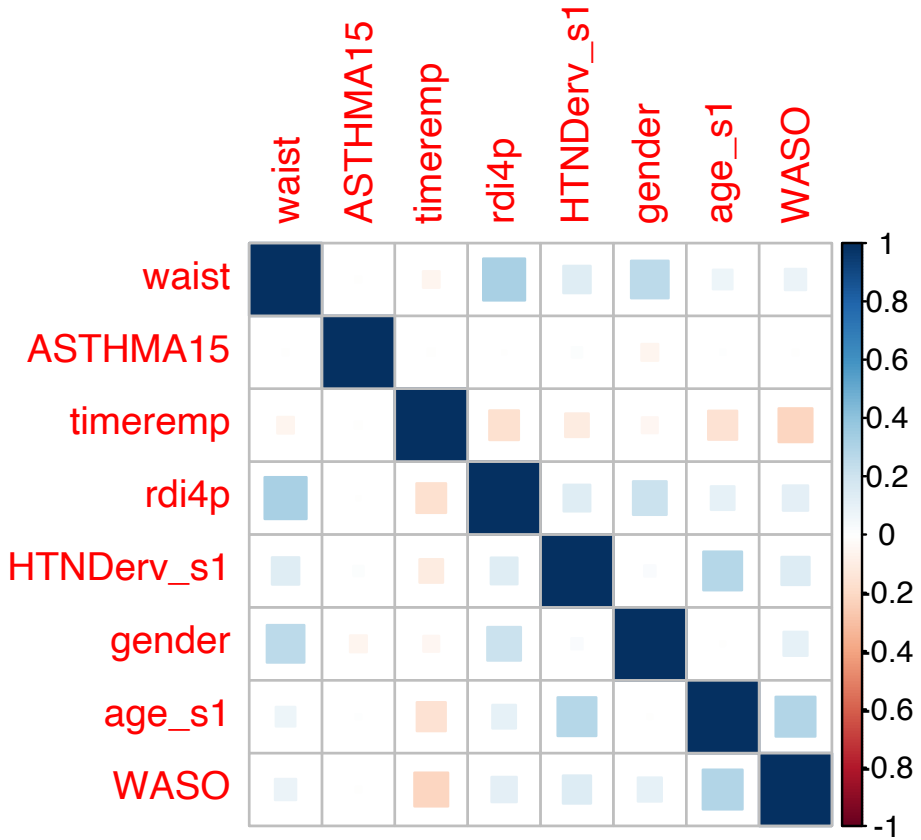
QQ-plots

- 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

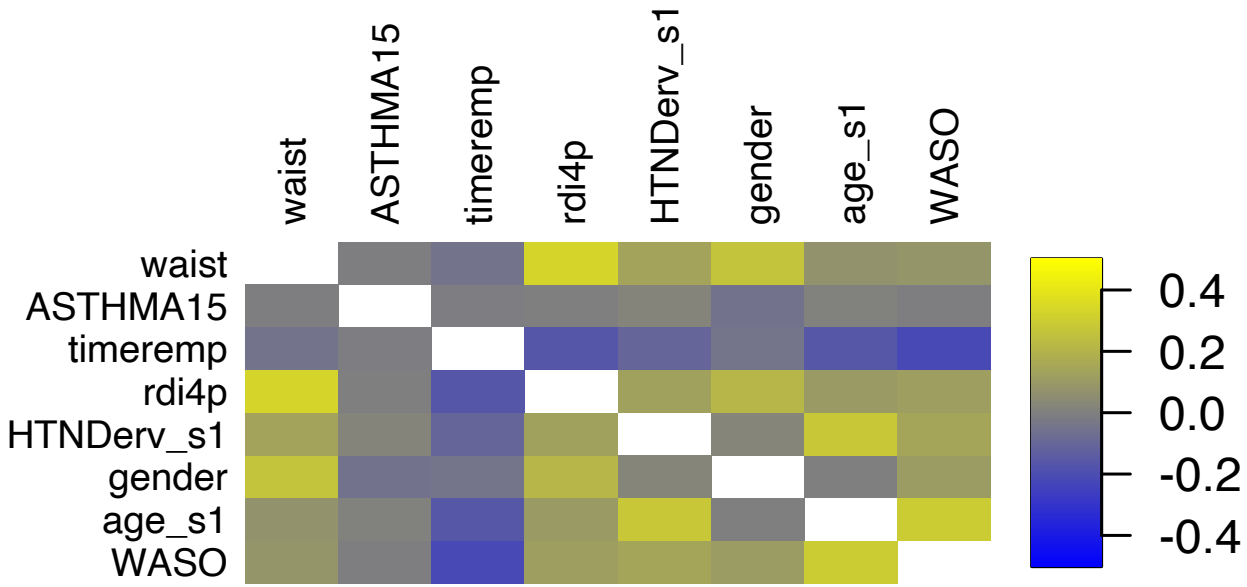
Heatmap of correlation

```
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")
```

Correlations among eight variables in SHHS



Correlations using the fields package



Heatmap all observations, standardized