WATERLOO | PHARMACY

PHARM 609
Advanced Pharmacokinetics

Winter 2016

Dagmar (Dasha) M. Hajducek
PhD, Statistics

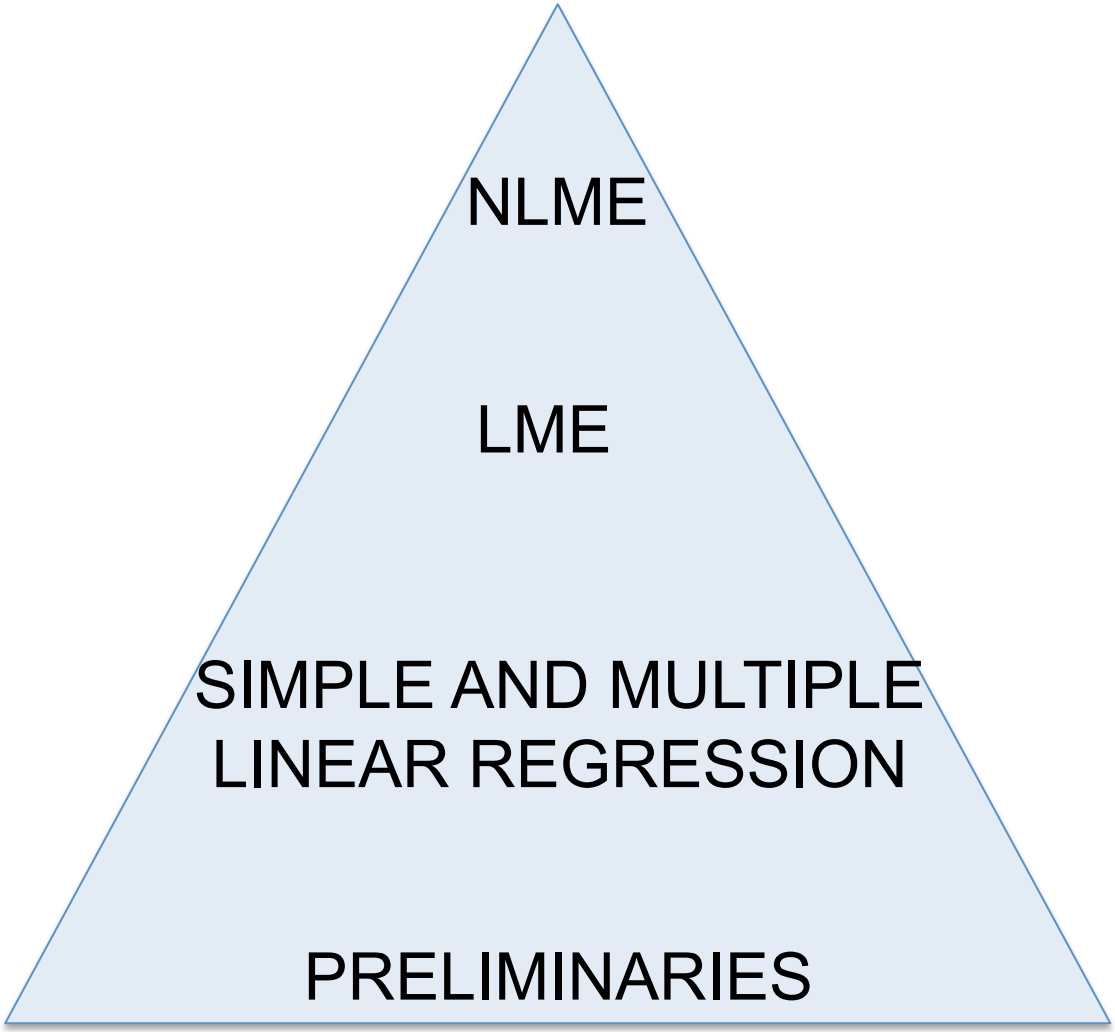
Course Description

- Provides statistical hands-on knowledge for the implementation of nonlinear mixed effects (NLME) models in the analysis of population pharmacokinetic data.

- Builds from simple to multiple linear regression models and to linear mixed effects models.

- Exploratory and descriptive analyses, as well as model implementation will be taught in R and Phoenix.

Course Slides



Overall Learning Objectives

Students will be able to:

1. Construct and interpret descriptive statistics and plots, basic continuous probability distributions, hypothesis tests and confidence intervals.
2. Perform simple and multiple linear regression modeling and interpret model diagnostics.
3. Perform linear mixed effects and non linear mixed effects modeling in the context of population pharmacokinetics and interpret model diagnostics.

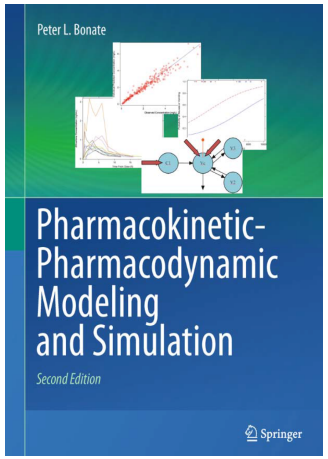
Selected Bibliography
Required reading

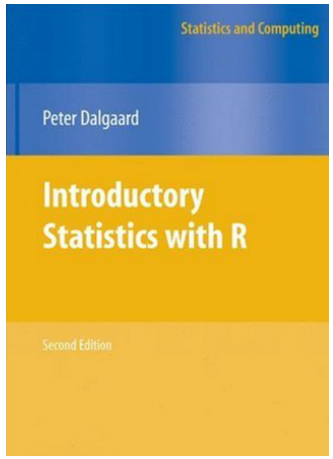
Available online @ UW Library Services

- Bonate, Peter L. *Pharmacokinetic-Pharmacodynamic Modeling and Simulation*. Springer, Second Edition, New York, 2011.
- Dalgaard, Peter. *Introductory Statistics with R.* Springer-Verlag, New York, 2nd Edition, 2008.

Linear & Multiple, LME, NLME

Preliminaries and throughout





Suggested Supplemental Bibliography

Preliminaries

- Durham, Todd A. and Turner, J. Rick. *Introduction to Statistics in Pharmaceutical Clinical Trials.* Pharmaceutical Press. 2008.
- Altman, Douglas G. *Practical Statistics for Medical Research.* Chapman & Hall/CRC, London, 1991. ***

Linear & multiple

- Montgomery, Douglas C., Peck, Elizabeth A., Vining Geoffrey, G. *Introduction to Linear Regression Analysis.* Wiley, 2012, 5th Edition.

LME, NLME

- Fitzmaurice, Garret M., Laird, Nan M., and Ware, James H. *Applied Longitudinal Analysis.* Wiley, 2nd Edition, 2011. ***
- Pinheiro, Jose C. and Bates, Douglas M. *Mixed Effects Models in S and S-Plus.* Springer-Verlag, New York, 2000.

*** To be added in syllabus.

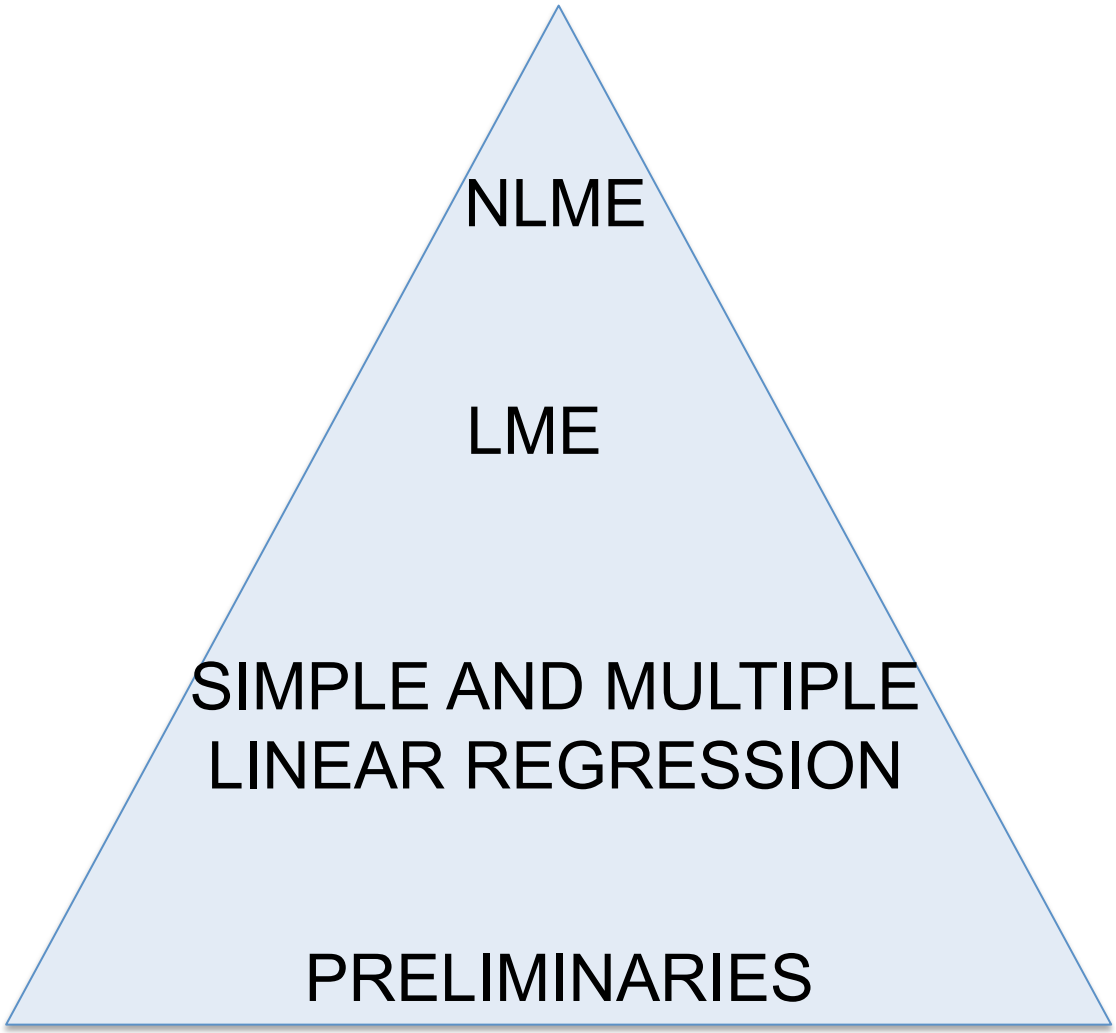
Selected Bibliography, Summarized

Bonate

Bonate,
Fitzmaurice (Supl)

Bonate,
Pinheiro (Supl)
Montgomery (Supl)

Dalgaard,
Durham, Altman (Supl)



Some online resources

About statistics

https://statswithcats.wordpress.com

Stats With Cats Blog

... for when you can't solve life's problems with statistics alone.

← 2014 in review

Share Your Career with Students →

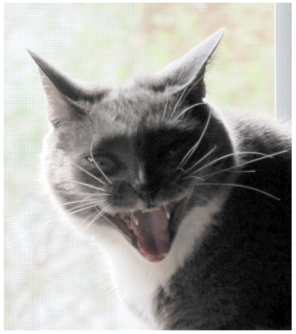
Search

Recent Posts

- It's Hard to be a Data-Driven Organization
- 1016
- Share Your Career with Students
- How to Tell if Correlation Implies Causation
- 2014 in review
- Reading Stats with Cats

How to Tell if Correlation Implies Causation

Posted on January 1, 2015



You've probably heard the admonition:

Correlation Does Not Imply Causation.

Everyone agrees that correlation is not the same as causation. However, those two words — correlation and causation — have generated quite a bit of discussion.

Why Causality Matters

About R

http://www.statmethods.net

https://cran.r-project.org/doc/contrib/Lemon-kickstart/

http://pj.freefaculty.org/R/Rtips.html (former StatsRUs)

http://www.math.csi.cuny.edu/Statistics/R/simpleR/ (pdf download)

Assignments

Assignment #1 (15%):
Use of R for descriptive analyses and plotting
Implementation of hypothesis testing.***

Assignment #2 (35%)
Implementation of simple and multiple linear regression

Assignment #3 (50%)
Implementation of LME*** and NLME for population pharmacokinetic analysis.

*** To be added in syllabus

Class:
Academic background?
Stats background?
R/RStudio installed?
R experience?

PRELIMINARIES

0. Introduction to R.
1. Statistical terms, descriptive statistics and plots.
2. Random variable, Normal, Lognormal, Sudent's t.
3. Sampling distribution and inferential statements under Normality.

0. Introduction to R

The R Software

- R is a statistical computer program.
- Provides an environment for statistical analysis and graphics.
- Available online under the General Public License (GPL) at

  http://www.r-project.org

- Recommended programming editor: RStudio

  https://www.rstudio.com

[Home]

Download

CRAN

R Project

About R
Contributors
What's New?
Mailing Lists
Bug Tracking
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

Documentation

Manuals
FAQs
The R Journal
Books
Certification
Other

Links

Bioconductor
Related Projects

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 3.2.3 (Wooden Christmas-Tree)** has been released on 2015-12-10.>
- **The R Journal Volume 7/1** is available.
- **R version 3.1.3 (Smooth Sidewalk)** has been released on 2015-03-09.
- **useR! 2015**, took place at the University of Aalborg, Denmark, June 30 - July 3, 2015.

RStudio

Products Resources Pricing About Us Blog

Welcome to RStudio - Open source and enterprise-ready professional software for R

Download RStudio | Discover Shiny | shinyapps.io Login

Powerful IDE for R

RStudio IDE is a powerful and productive user interface for R. It's free and open source, and works great on Windows, Mac, and Linux.

Learn More ›

R Packages

Our developers and expert trainers are the authors of several popular R packages, including ggplot2, plyr, lubridate, and others.

Learn More ›

Bring R to the web

Shiny is an elegant and powerful web framework for building interactive reports and visualizations using R — with or without web development skills.

Learn More ›

RStudio Running…

RStudio

CysticFibrosis.R

```
1  setwd("~/Google Drive/Teaching/R Codes/")
2
3
4  #jpeg("~/Google Drive/Teaching/COURSES/Pictures/hist5FU.jpg")
5
6  # require(ISwR)
7  library(ISwR)
8
9  ?cystfibr
10
11 dat <- cystfibr
12 names(dat)
13
14
15 # ----------------------------
16 # correlation
17 # ----------------------------
18 # correlation matrix for pemax and nine potential explanatory variables
19 round(cor(dat),3)
20
21 # scatter diagrams
22 par(mex=0.5)
23 #jpeg("~/Google Drive/Teaching/COURSES/Assessments/Multiple/scatter.jpg")
24 pairs(dat, upper.panel=NULL, gap=0, cex.labels=0.9)
25 #graphics.off()
26
```

25:1 (Untitled) R Script

Console ~/Google Drive/Teaching/R Codes/

```
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> setwd("~/Google Drive/Teaching/R Codes/")
> library(ISwR)
> dat <- cystfibr
> names(dat)
 [1] "age"    "sex"    "height" "weight" "bmp"    "fev1"   "rv"     "frc"    "tlc"    "pemax"
> par(mex=0.5)
> pairs(dat, upper.panel=NULL, gap=0, cex.labels=0.9)
>
```

Environment History

Import Dataset Clear List

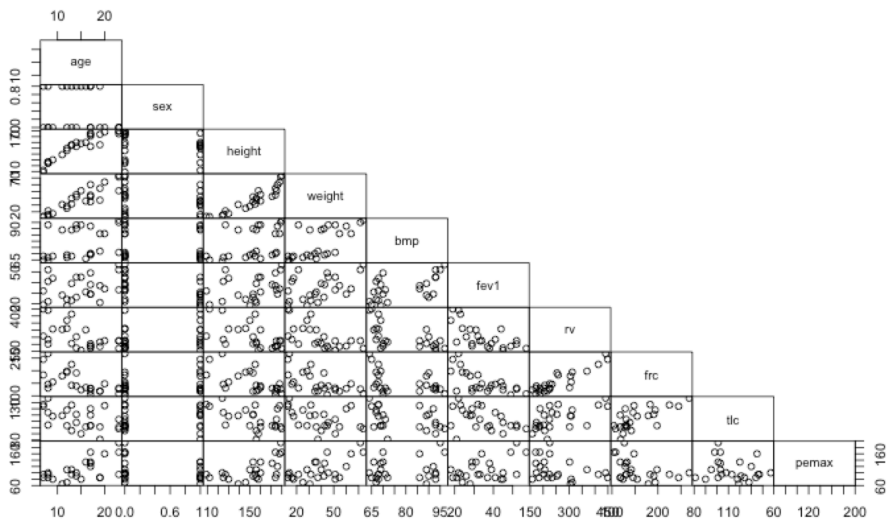
Global Environment

Data

a	1 obs. of 5 variables
b	1 obs. of 5 variables
c	1 obs. of 5 variables
cosa	354 obs. of 3 variables
cosa1	97 obs. of 2 variables
cov12	num [1, 1] 27.8
cov12.2	num [1:11, 1:11] 51.4 -23.3 -23.6 -23.3 -23.3 ...
d	1 obs. of 5 variables
dat	25 obs. of 10 variables
dat.diffs	354 obs. of 2 variables
dat.pcs	354 obs. of 2 variables

Files Plots Packages Help Viewer

Zoom Export Clear All



Some general characteristics of R:

- Case-sensitive, interpreted language (compiled languages such as C are faster).
- Handles a variety of data types:
  - Vectors: numerical, character and logical
  - Data frames
  - Matrices
  - Lists

  } Mostly used here

- Works with built-in and user-created functions
- Basic functions are available by default. Others are contained in packages online for free.

Getting Started in R

```
Basic arithmetic operations:        +   -   /   * ^   ()
# symbol = commenting out
```

Can also be done through the drop-down menus in RStudio!

```
# Getting the working directory- Session menu in RStudio
getwd()

# Setting the working directory – Session menu in RStudio
setwd()
e.g. setwd("~/PHARM609/R Codes/")

# Installing packages
# after selecting the mirror site to download, the package
# "lattice" will be installed.
install.packages("lattice")

# Getting help
?mean          # for the function called "mean"
??plot         # will search all the libraries and functions
               #containing the word "plot"
```

Importing and exporting data files

```
# Reading a .csv file
mydata <- read.csv("c:\\...\\data.csv",header=T)


# Reading a .txt file (space, tab or comma separated)
mydata <- read.table("c:\\...\\data.txt",sep="",header=T)

# sep="" implies the data in a row is separated by white
# space, which is default.


# Export data
write.table(mydata,"c"\\...\\data.txt",sep="")
```

Accessing Built in Data

1. Install the R package that contains the data set.

Some data examples here are taken from Dalgaard's book, available in the R package:

"ISwR" (as in Introductory Statistics with R, by Peter Dalgaard)

Can select the "Install Packages" from the "Tools" menu in Rstudio.

Documentation of the packages can be found online.
E.g. https://cran.r-project.org/web/packages/ISwR/ISwR.pdf

2. Load the library of the R package. E.g., by typing `library(ISwR)`

3. Call the help for the data set of interest, which will give you a description. E.g. `?cystfibr`

Example: Maximal Static Expiratory Pressure (Pemax, cm H_20) in cystic fibrosis patients as a measure of respiratory strength (Dalgaard, 2008).

```
> # first need to install the ISwR package (Tools option in RStudio)
> library(ISwR) # loading the library associated with ISwR
> ?cystfibr
```

Description

The `cystfibr` data frame has 25 rows and 10 columns. It contains lung function data for cystic fibrosis patients (7–23 years old).

Usage

```
cystfibr
```

Format

This data frame contains the following columns:

`age`

a numeric vector, age in years.

`sex`

a numeric vector code, 0: male, 1:female.

...

Source

D.G. Altman (1991), *Practical Statistics for Medical Research*, Table 12.11, Chapman & Hall.

References

O'Neill et al. (1983), The effects of chronic hyperinflation, nutritional status, and posture on respiratory muscle strength in cystic fibrosis, *Am. Rev. Respir. Dis.*, 128:1051–1054.

Cystic Fibrosis Data Example

```
# getting the names of the variables
> names(cystfibr)
 [1] "age"    "sex"    "height" "weight" "bmp"    "fev1"   "rv"
"frc"    "tlc"    "pemax"


# getting the structure of the data set
> str(cystfibr)
'data.frame':25 obs. of  10 variables:
 $ age   : int  7 7 8 8 8 9 11 12 12 13 ...
 $ sex   : int  0 1 0 1 0 0 1 1 0 1 ...
 $ height: int  109 112 124 125 127 130 139 150 146 155 ...
 $ weight: num  13.1 12.9 14.1 16.2 21.5 17.5 30.7 28.4 25.1 ..
 $ bmp   : int  68 65 64 67 93 68 89 69 67 68 ...
 $ fev1  : int  32 19 22 41 52 44 28 18 24 23 ...
 $ rv    : int  258 449 441 234 202 308 305 369 312 413 ...
 $ frc   : int  183 245 268 146 131 155 179 198 194 225 ...
 $ tlc   : int  137 134 147 124 104 118 119 103 128 136 ...
 $ pemax : int  95 85 100 85 95 80 65 110 70 95 ...
```

Basic commands with data frames
Viewing portions of data

```
> head(cystfibr) # view the top
  age sex height weight bmp fev1  rv frc tlc pemax
1   7   0    109   13.1  68   32 258 183 137    95
2   7   1    112   12.9  65   19 449 245 134    85
3   8   0    124   14.1  64   22 441 268 147   100
4   8   1    125   16.2  67   41 234 146 124    85
5   8   0    127   21.5  93   52 202 131 104    95
6   9   0    130   17.5  68   44 308 155 118    80
> tail(cystfibr) # view the bottom
   age sex height weight bmp fev1  rv frc tlc pemax
20  19   1    156   37.2  72   21 216 119  81    85
21  19   0    174   54.6  86   37 184 118 101    85
22  20   0    178   64.0  86   34 225 148 135   160
23  23   0    180   73.8  97   57 171 108  98   165
24  23   0    175   51.1  71   33 224 131 113    95
25  23   0    179   71.5  95   52 225 127 101   195

> cystfibr$age  # view the age variable
 [1]  7  7  8  8  8  9 11 12 12 13 13 14 14 15 16 17 17 17
17 19 19 20 23 23 23
```

Basic commands with data frames
Data subsets with $ symbol

```
dataset[ rows , columns ]
```

```
> # select the 10th row in cystfibr data
> cystfibr[10,]
   age sex height weight bmp fev1  rv frc tlc pemax
10  13   1    155   31.5  68   23 413 225 136    95

> # select subjects with age <10 yrs
> cystfibr[cystfibr$age<10,]
  age sex height weight bmp fev1  rv frc tlc pemax
1   7   0    109   13.1  68   32 258 183 137    95
2   7   1    112   12.9  65   19 449 245 134    85
3   8   0    124   14.1  64   22 441 268 147   100
4   8   1    125   16.2  67   41 234 146 124    85
5   8   0    127   21.5  93   52 202 131 104    95
6   9   0    130   17.5  68   44 308 155 118    80
```

Dalgaard calls the use of $ clumsy

Basic commands with data frames
Data subsets with attach() , detach()

- attach() separates the columns of the data set so that they can be accessed by simply giving their names.

- detach() reverses the action.

Dalgaard prefers attach(), detach() to $

```
> attach(cystfibr)
> age
 [1]  7  7  8  8  8  9 11 12 12 13 13 14 14 …
> cystfibr[age<10,]
  age sex height weight bmp fev1  rv frc tlc pemax
1   7   0    109   13.1  68   32 258 183 137    95
2   7   1    112   12.9  65   19 449 245 134    85
3   8   0    124   14.1  64   22 441 268 147   100
4   8   1    125   16.2  67   41 234 146 124    85
5   8   0    127   21.5  93   52 202 131 104    95
6   9   0    130   17.5  68   44 308 155 118    80
> detach()
> age
Error: object 'age' not found
```

After detach() the age variable is no longer available directly in the workspace

Basic commands with data frames
Data subsets with subset()

```
> subset(cystfibr,age<10)
  age sex height weight bmp fev1  rv frc tlc pemax
1   7   0    109   13.1  68   32 258 183 137    95
2   7   1    112   12.9  65   19 449 245 134    85
3   8   0    124   14.1  64   22 441 268 147   100
4   8   1    125   16.2  67   41 234 146 124    85
5   8   0    127   21.5  93   52 202 131 104    95
6   9   0    130   17.5  68   44 308 155 118    80
```

```
> # calculate the mean for each variable in the data
set
> lapply(cystfibr,mean)
$age
[1] 14.48
$sex
[1] 0.44
$height
[1] 152.8
$weight
[1] 38.404
$bmp
[1] 78.28
$fev1
[1] 34.72
…
```

Note the sex variable is coded as 1's and 0's so this mean of 0.44 doesn't make sense here.

Creating vectors c(), seq(), rep(), :

```
> # creating vectors
> myvector <- c(4,6,9,12)
> myvector
[1]  4  6  9 12

> # creating a vector using a
sequence
> myseq1 <- seq(1,5,.5)
> myseq1
[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0
4.5 5.0

> # creating a vector using colon
punctuation
> myseq2 <- 1:5
> myseq2
[1] 1 2 3 4 5

> # getting the length of a
vector
> length(myseq1)
[1] 9
```

```
> # generating repeated values
> id <- rep(c(1,2,3),each=2)
> id
[1] 1 1 2 2 3 3
```

Binding columns and rows as vectors cbind() & rbind()

```
> # binding columns in a matrix
> mymat1 <- cbind(X=1:4,Y=c(1,3,7,5),Z=rep(1,4))
> mymat1[,1] # accessing the X column
[1] 1 2 3 4
> mymat1[1,]
X Y Z
1 1 1
```

mymat1 and mymat2 are matrices in R and can be used for matrix algebra operations (not really needed here).

```
> # binding rows in a matrix
> mymat2 <- rbind(X=1:4,Y=c(1,3,7,5),Z=rep(1,4))
> mymat2
  [,1] [,2] [,3] [,4]
X    1    2    3    4
Y    1    3    7    5
Z    1    1    1    1
```

Data frames may be constructed using the data.frame() command.

Data Frames

```
> # creating a data frame
> mydata.frame <- data.frame(X=1:4,Y=c(1,3,7,5),Z=rep(2,4))
> mydata.frame
  X Y Z
1 1 1 2
2 2 3 2
3 3 7 2
4 4 5 2
> mydata.frame$X          # accessing column X
[1] 1 2 3 4
> mydata.frame[,1]        # accessing column X
[1] 1 2 3 4
> mydata.frame[mydata.frame$X==1,]     # accessing the row(s)
  X Y Z                                where X=1
1 1 1 2
> mydata.frame[mydata.frame$X<3,]      # accessing the row(s)
  X Y Z                                where X<3
1 1 1 2
2 2 3 2
```

An advantage of a data frame over a matrix is that variables and rows can be subsetted as shown earlier, using the $ sign or by using the attach()& detach commands.

Exercise

1. From the `cystfibr` data set, subset the rows corresponding to individuals: (a) with values greater than 44 of forced expiratory volume and (b) maximum expiratory pressure equal to 165.

Hint: type `?cystfibr` to see the names of the variables.
Hint: logical equality is expressed with "==" and the symbol "&" can be used to impose an additional logical statement. E.g."x>15 & y==9"

Use the three options below:
the `subset()` function,
the `attach()` function,
the `$` symbol.

2. (a) Construct a data frame with the variables age, sex, bmp and fev1 of the cystfibr data, using data frame() function.
(b) Change the spelling of this variable names to capital letters.
(c) Convert the sex variable originally coded as numeric to a factor with levels "females" and "males" – hint: use the `factor()` function as in `factor(variable, levels=c("name1","name2")`

The R Workspace

The place where all created data sets and variables are stored.

The list function will give the variables or data sets in the workspace
the remove function will delete the variables or data sets that you specify:

```
ls(), rm(var0iable1,variable2),
```

`rm(list=ls())` removes all variables from the workspace.

The R Environment

The R Environment is the place that stores:

- The R Workspace
- All the files necessary for running the R Program: built-in and installed by the user.

The R Environment can be viewed through the search function. The following 9 files are shown by default:

```
> search()
[1] ".GlobalEnv" "package:stats" "package:graphics"
[4] "package:grDevices" "package:utils" "package:datasets"
[7] "package:methods" "Autoloads" "package:base"
```

Files are sorted, and the first file corresponds to the R Workspace. The functions attach() and detach() will create an additional file.

```
> attach(cystfibr)
The following object is masked from package:ISwR:

    tlc

> search()
> search()
[1] ".GlobalEnv" ”cystfibr" "package:ISwR"
[4] "package:stats" "package:graphics" "package:grDevices"
[7] "package:utils" "package:datasets" "package:methods"
[10] "Autoloads" "package:base”
```

- “masked” above means that there is one object (called “tlc”) in the ISwR package with variables of the same name as in the cystfibr data set (such as age, weight, height).

- R looks for the variables in the environment in the same order of the files
  - E.g. since “cystfibr” appears as #2 and the “package:ISwR” is the #3, every time you access age, weight or height, R will use those stored in “cystfibr”.

- The R Environment is also accessible in RStudio, by clicking on the “Environment” tab in the upper right window.

1. Statistical terms, descriptive statistics and plots

Specific Learning objectives:

1.1. Explain the following terms:
Units, population, sample,
Parameter, statistic,
Variable, types of variables, observations.

1.2. Interpret, calculate or plot the following:
Measures of spread
Measures of location
Boxplots, histograms, QQ Plots.

Statistics, Definition and Role in PK/PD

Statistics is the science of **collection**, **analysis** and **interpretation** of data providing a ground for educated decisions in the face of **variability** and **uncertainty**.

“The best thing of being a statistician is you get to play in everyone else’s back yard”

John Tukey, Bell Labs, Princeton University

Statistics' Role in PK/PD

In its simple sense, pharmacometrics is the science of quantitative pharmacology, but has been more formally defined as:

"the science of developing and applying mathematical and statistical methods to characterize, understand and predict a drug's pharmacokinetic, pharmacodynamic and biomarker-outcomes behavior" (Williams & Ette 2010).

Bonate, 2011.

Units, population, sample

- Units: subjects or objects under study.
  - E.g. a person under a specific medical condition, a healthy individual, a Petri dish with cultured cells.

- Population (target): collection of all units which share a common characteristic of research interest, to which the study findings are generalized.
  - E.g. in a drug trial: the group of people researchers think might benefit from a particular drug.

- Sample: subset of the population used to infer on the population's characteristics
  - E.g. in a drug trial, a sample is taken from the population to infer about a quantity of interest.

Parameter, Statistic

- Parameter: unknown population characteristic.
  - E.g. population mean clearance, regression model coefficient, population variance.

- Statistic: sample characteristic, a function of the data.

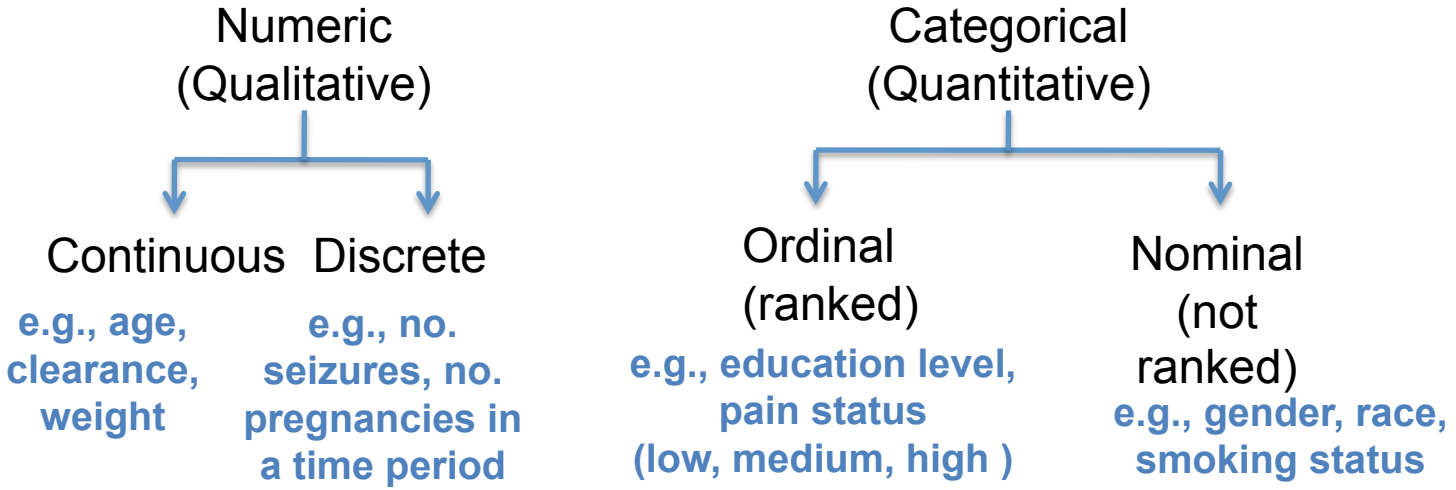
Some statistics are used for estimation of parameters, others for hypothesis testing.

  - E.g. sample mean clearance, estimated regression model coefficient, sample variance, test statistics.

Variables, Types of variables, Observations

- Variables: characteristics of each population unit, can be observed or manipulated.
  - Response variable: primary variable of interest.
  - Covariates: other variables attached to each unit.

Types of variables: **why is this important?**



Observation: a realization, numerical/categorical value taken by a variable.

Descriptive statistics

- Measures of location: quantities that describe portions of the data.

  Median, mean : center of the data, aka *"measures of central tendency"*

  Quartiles : fractions of the data

- Measures of spread: quantities that describe the variability in the data.

  Sample variance, standard deviation, coefficient of variation, range, interquartile range.

- Descriptive plots: graphical summaries of the distribution of the data

  Histograms, cumulative histograms, box plots, quantile-quantile plots.

Descriptive statistics: Measures of Location

Median: value that comes half-way when data are ranked in order.
For an even number of observations it is the average of the two central values.

Quartiles: divide data in four parts (quarters):

Q1 : First quartile, 25th percentile

Q2 : Second quartile, 50th percentile

Q3 : Third quartile, 75th percentile

Median of the upper 50% portion of the data

Median of all the data

Median of the lower 50% portion of the data

Mean: sum of observations divided by the number (n) of observations:

$$\overline{X} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The mean is sensitive to extreme observations in which case the median is preferred.

Example: Hypothetical Serum Albumin data from patients with Primary Biliary Cirrhosis (PBC).

Based on Altman, 1990.

```
>dat <- read.csv("Data/SerumAlbuminSimulated.csv")

> head(dat)
  SerumAlbumin    Sex
1       22.275 female
2       44.163 female
3       49.955 female
4       39.854 female
5       32.923 female
6       33.195 female

> str(dat)
'data.frame':  600 obs. of  2 variables:
 $ SerumAlbumin: num  22.3 44.2 50 39.9 32.9 ...
 $ Sex         : Factor w/ 2 levels "female","male": 1 1 1 1 1 1. . .

> table(dat$Sex)
female   male
   300    300
```

table() counts the number of observations by gender

Example: Hypothetical Serum Albumin data from patients with Primary Biliary Cirrhosis (PBC). Based on Altman, 1990.

```
> attach(dat)
> SA.female <- SerumAlbumin[Sex=="female"]
> SA.male <- SerumAlbumin[Sex=="male"]
> detach()


> n.SAfemale <- length(SA.female)
> mean.SAfemale <- mean(SA.female)
> quarts.SAfemale <- quantile(SA.female,c(.25,.5,.75))


> n.SAfemale
[1] 300
> mean.SAfemale
[1] 40.14545
> quarts.SAfemale
     25%      50%      75%
32.72200 41.22050 46.18675

> can also try the summary() function
> summary(SA.female)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.30   32.72   41.22   40.15   46.19   62.08
```

Descriptive statistics: Measures of spread

Sample variance: "Corrected average deviation from the mean"

Individual deviations from the mean

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Average: n summands divided by (n-1), where (n-1) is used as a correction ("Bessel's correction")

Sample standard deviation: the square root of s^2 (back to the original scale of the data).

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

Descriptive statistics: Measures of spread

Coefficient of variation: measure of relative dispersion. $CV = \frac{s}{\overline{X}}$

Range: measure of dispersion based on subtracting the minimum to the maximum data value:

$$Range = Max - Min$$

Interquartile Range: measure of dispersion based on subtracting the first quartile (Q1) from the third (Q3):

$$IQR = Q3 - Q1$$

Looking at plots later on will make more intuitive sense of these.

Example: Hypothetical Serum Albumin data from patients with Primary Biliary Cirrhosis (PBC). Based on Altman, 1990.

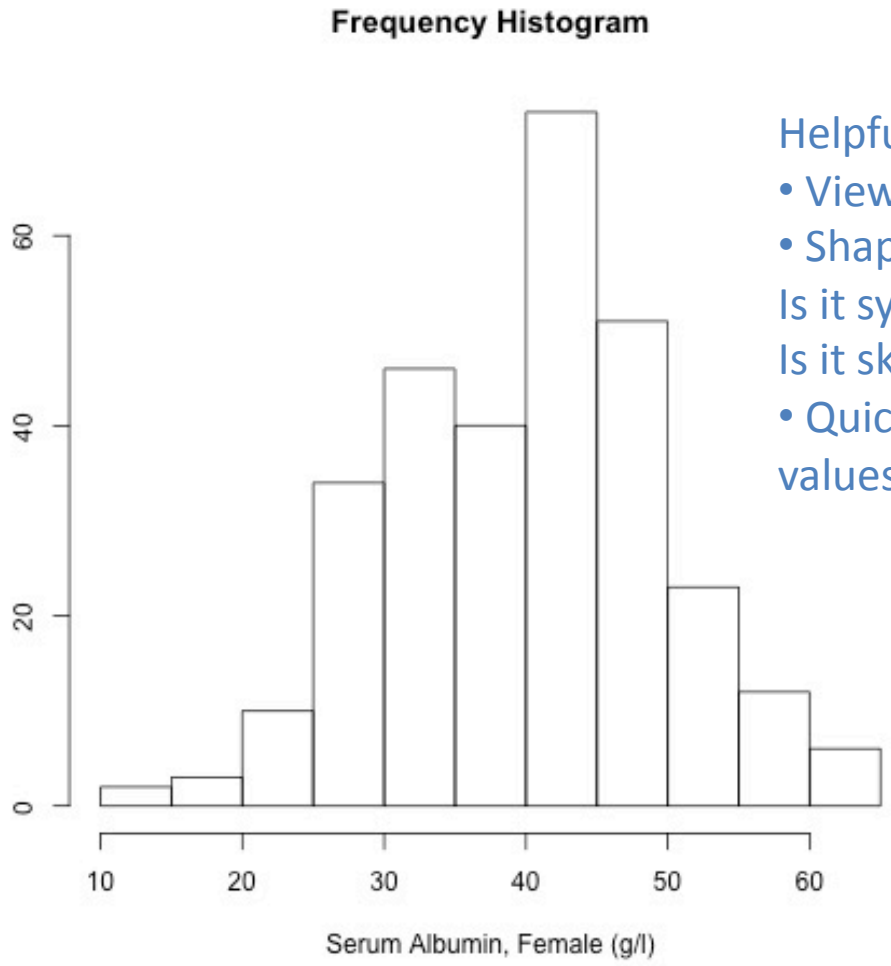
```
> var.SAfemale <- var(SA.female)
> sd.SAfemale <- sd(SA.female)
> range.SAfemale <- range(SA.female)[2]-range(SA.female)[1]
> cv.SAfemale <- sd.SAfemale / mean.SAfemale
> IQR.SAfemale <- quarts.SAfemale[3]-quarts.SAfemale[1]
```

```
> var.SAfemale
[1] 89.86755

> sd.SAfemale
[1] 9.47985

> cv.SAfemale
[1] 0.2361376

> range.SAfemale
[1] 48.785

> IQR.SAfemale
     75%
13.46475
```

Basic descriptive plots: Histogram

- Histogram: plot of the frequency distribution of a continuous variable.

Y-Axis=Frequency, e.g., Number of subjects



Helpful for
- Viewing extreme observations
- Shape of the distribution:

Is it symmetric?
Is it skewed?
- Quick look at central tendency values and overall spread

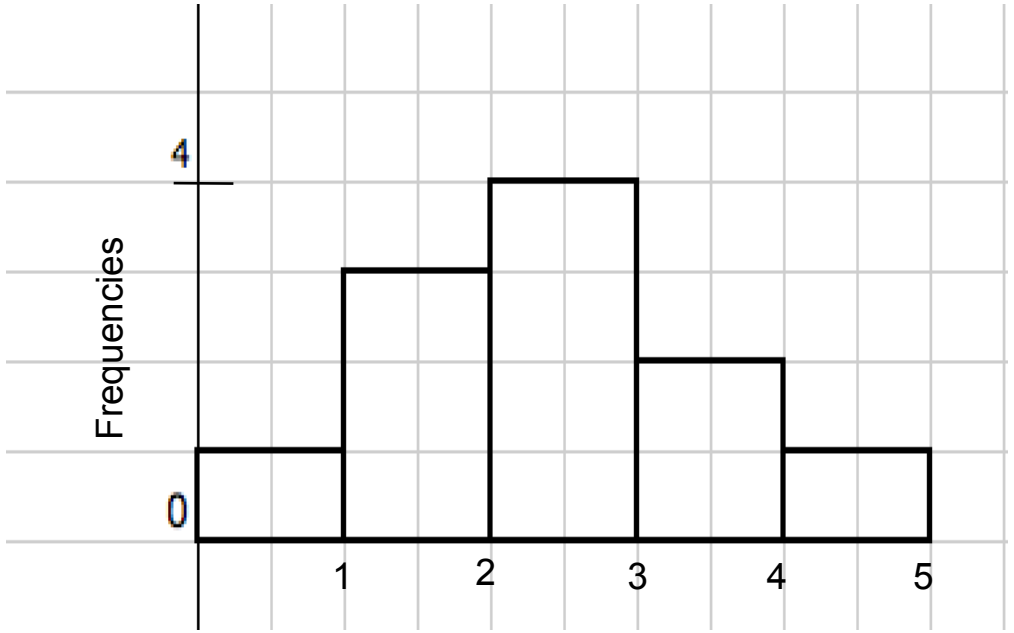
```
hist(SA.female,main="Frequency Histogram")
```

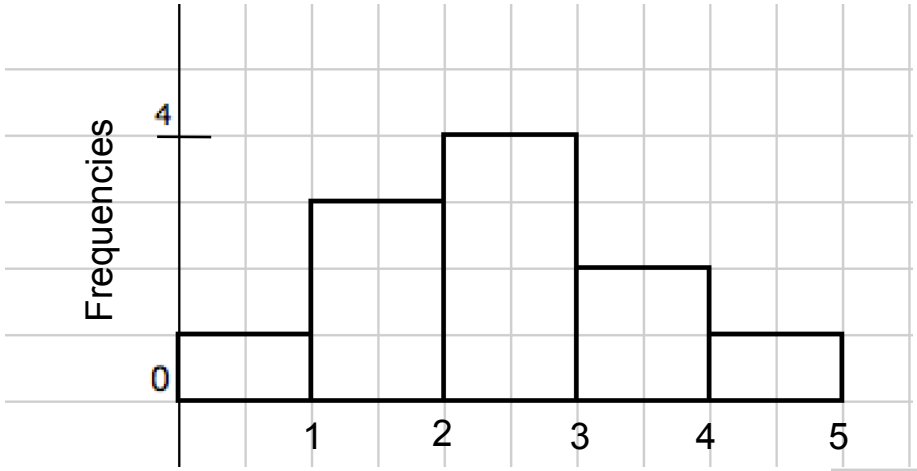
Visualizing Construction of a Cumulative Histogram

Imagine a data set with the following values:

1 2 2 2 3 3 3 3 4 4 5

A histogram may look like this:





Visualizing Construction of a Cumulative Histogram

1. Stack the 1st bin in the histogram (representing frequency=1) on top of the next bin to the right, and so on…

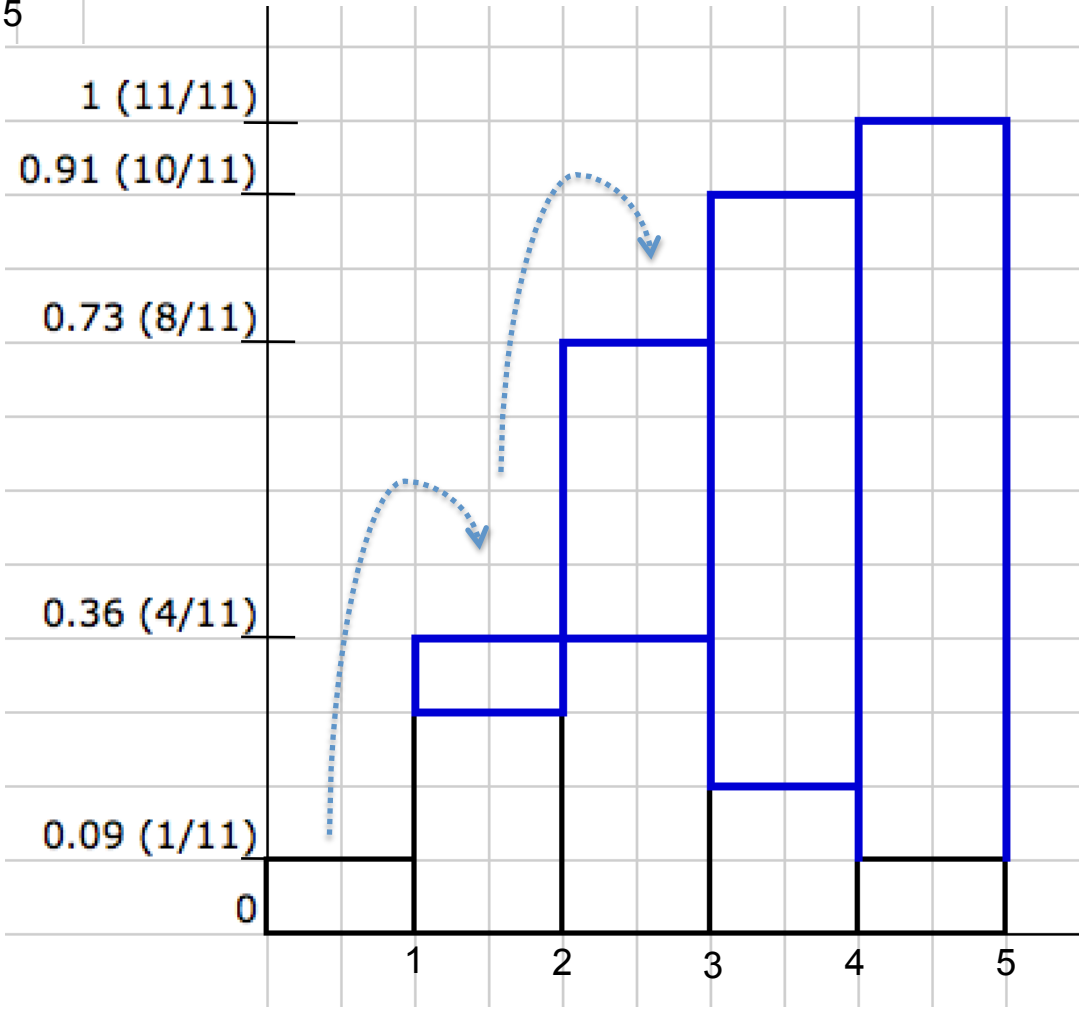
2. Re-scale the frequencies axis to go from 0 to 1 by dividing each y-axis value over the maximum cumulative frequency (11).

Proportion of sample units below or equal to

2 is 0.36,

3 is 0.73,….

Cumulative sums of frequencies:



Basic descriptive plots: Cumulative Relative Histogram

- Cumulative histogram: plot of the cumulative relative frequencies.

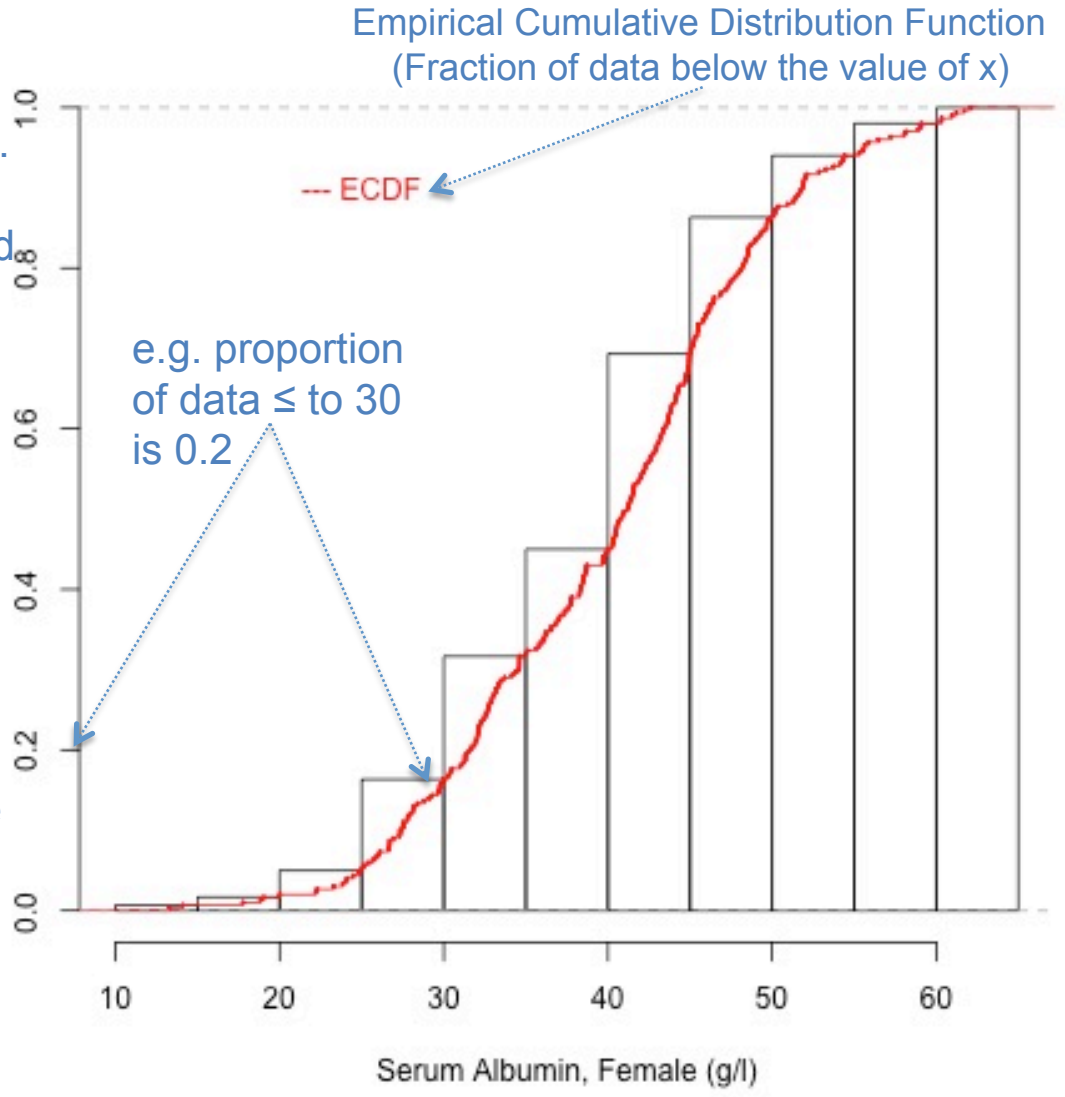
Y-Axis=Cumulative Relative Frequencies, i.e., sum of proportions.

- Note it always goes from 0 to 1.
- Can be considered a sample based probability:

e.g., Pr(Serum Albumin ≤ 30) = 0.2

This plot can be used to compare with the shape of a theoretical probability distribution, e.g., Normal, logNormal, etc.

When using cumulative frequencies, the scale of the y-axis changes to the actual frequencies.



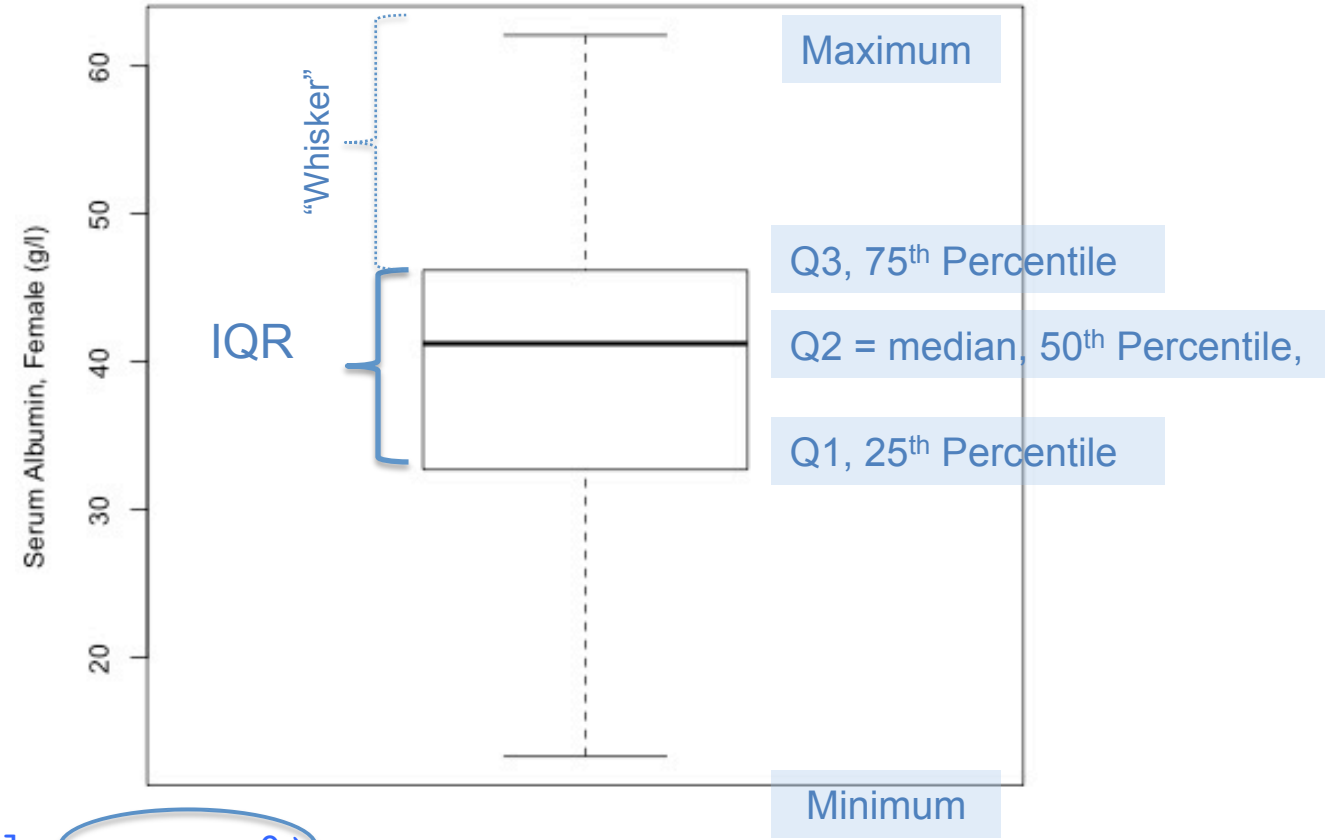
Cumulative Relative Histogram in R
In class exercise

```
> h=hist(SA.female)
> h$counts
 [1]  2  3 10 34 46 40 73 51 23 12  6

> cumsum(h$counts)
 [1]   2   5  15  49  95 135 208 259 282 294 300

> m.cum <- max(cumsum(h$counts))

# replace the relative cell freqs by cumulative relative freqs:
> h$counts <- cumsum(h$counts)/m.cum
> h$counts
 [1] 0.006666667 0.016666667 0.050000000 0.163333333 0.316666667 0.450000000
 [7] 0.693333333 0.863333333 0.940000000 0.980000000 1.000000000

# plot a cumulative relative histogram of y:
plot( h, ylim=c(0,1),xlab="Serum Albumin, Females (m/l)",
                              main="Cumulative Histogram")
lines(ecdf(x),cex=.2,col="red")
text(25,.9,"--- ECDF",col="red")
```

Basic descriptive plots: Box Plots (Box and Whisker Plots)

Used to view:

- The location and spread of the distribution
- Symmetry of the distribution
- How typical values are concentrated in relation to the spread (e.g. Interquartile Range: IQR= Q3 – Q1)

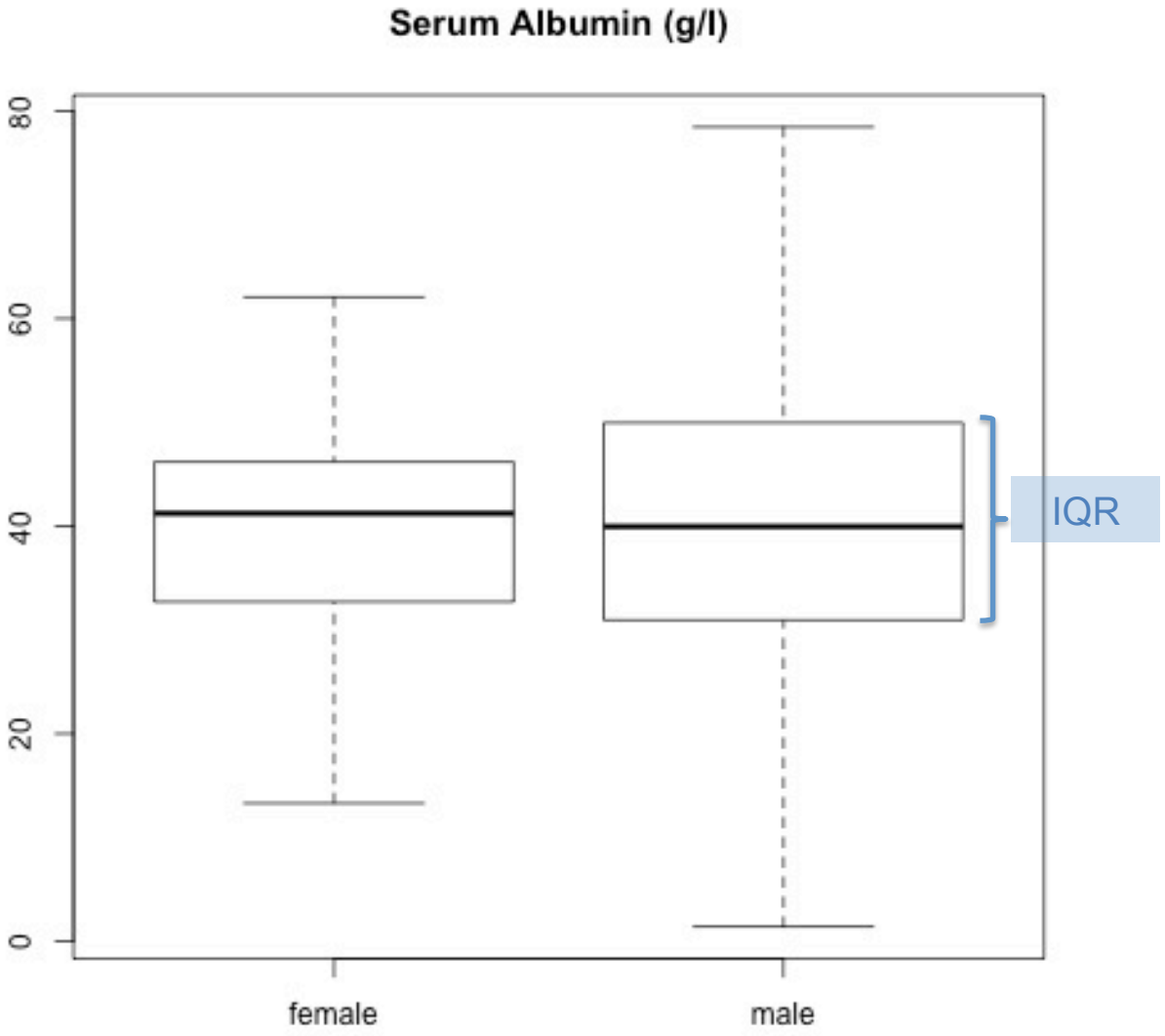



```
boxplot(Variable,range=0)
```

Comparative Box Plots

E.g., Serum Albumin patients with Primary Biliary Cirrhosis (PBC) in females and males

Median values are close, however,
males have a larger variability as whiskers are longer and interquartile range (IQR=Q3-Q1) is wider.




```
boxplot(SerumAlbumin~Sex,data=dataset,range=0,main="Serum Albumin (g/l)")
```

`boxplot()` and `range`

Detecting potential outliers by setting the `range` value

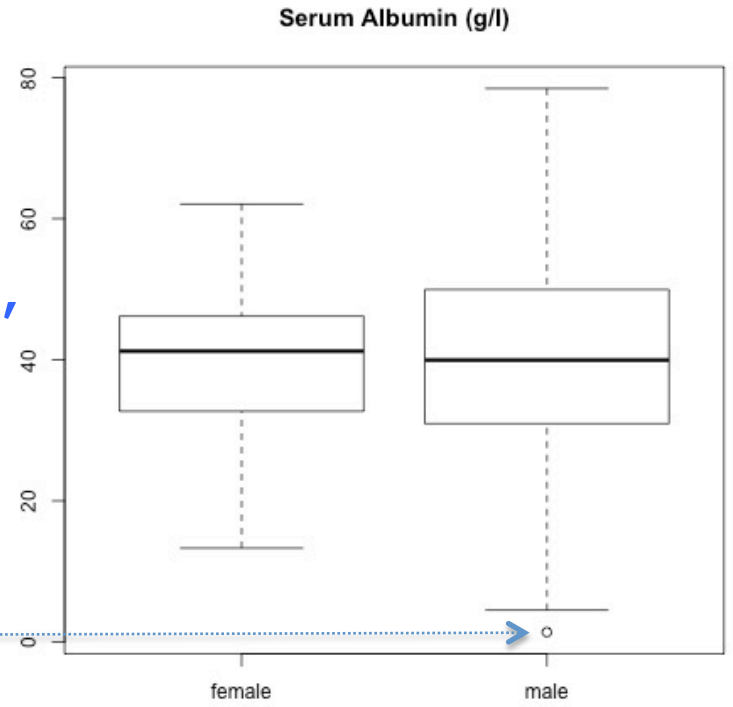
- Setting `range=0` will plot the whiskers up and down to the max and min of the data.

```
boxplot(SerumAlbumin~Sex, range=0, data=dataset)
```

- Setting e.g. `range=1.5` or excluding it will plot the whiskers:
  - up to 1.5xIQR above Q3 and
  - down to 1.5xIQR below Q1:

```
boxplot(SerumAlbumin~Sex,range=1.5,
                data=dataset)
boxplot(SerumAlbumin~Sex,
                data=dataset)
```

Potential outlier



Basic descriptive plots: Quantile-Quantile Plots

Quantiles are defined the same as percentiles but are referred in terms of sample fractions instead of percentages.

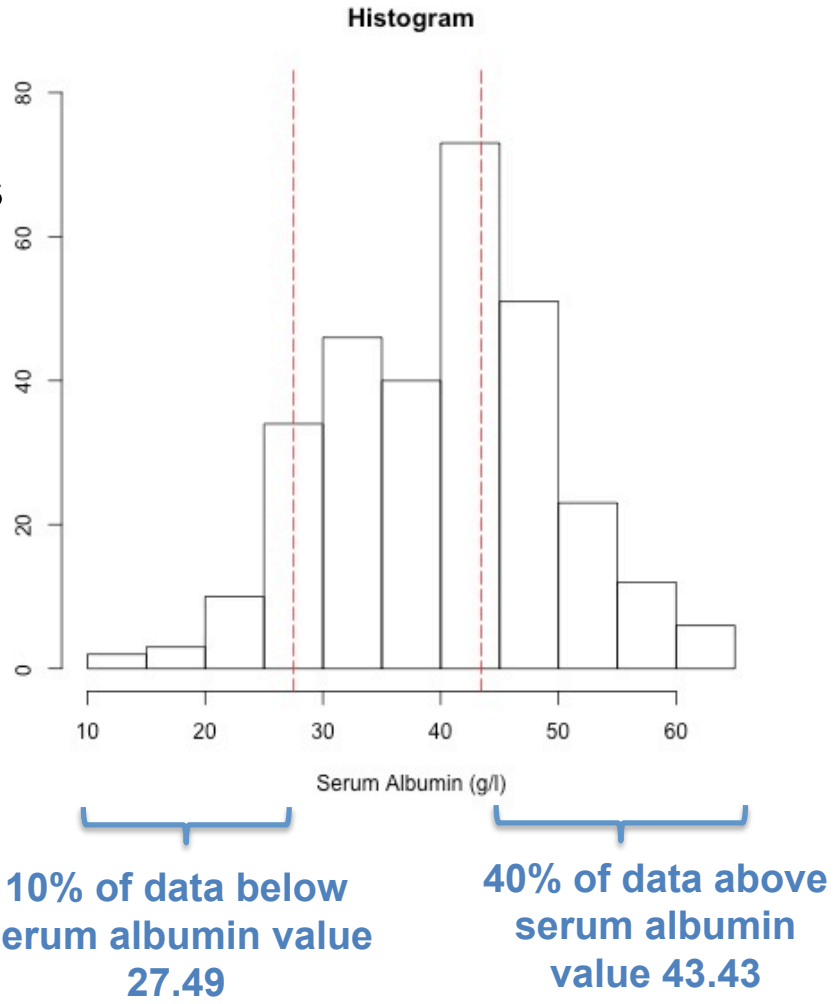
Types of Quantiles

Quartiles: divide data into quarters
Quintiles: divide data into fifths
Deciles: divide data into tenths

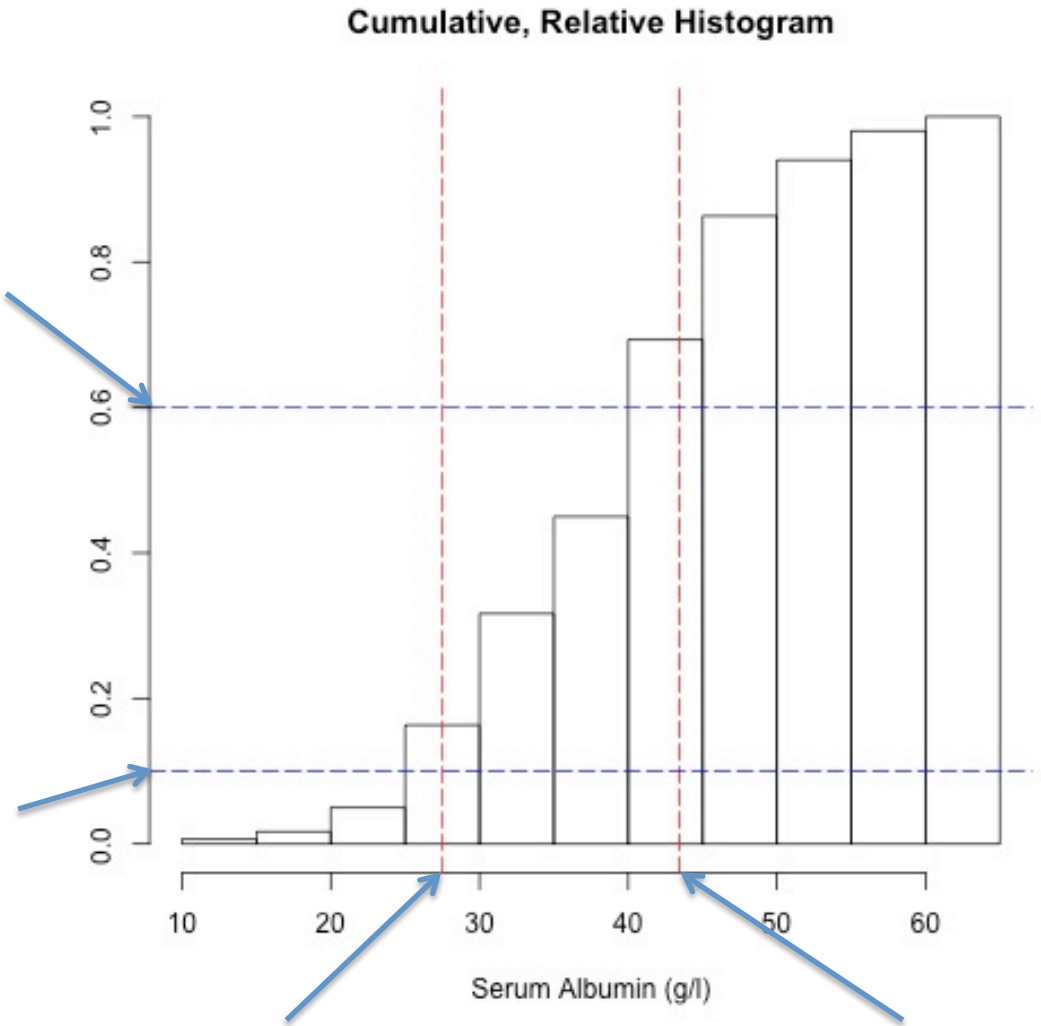
Also called 1st decile

E.g.

- Quantile corresponding to 10th percentile: 27.49
- Quantile corresponding to 60th percentile=43.43



Same quantiles shown on a Cumulative Histogram:



Cumulative relative frequency = 0.60 (associated with 60th percentile)

Cumulative relative frequency = .10 (associated with 10th percentile)

27.49 g/l of Serum Albumin (quantile corresponding to 10th percentile)

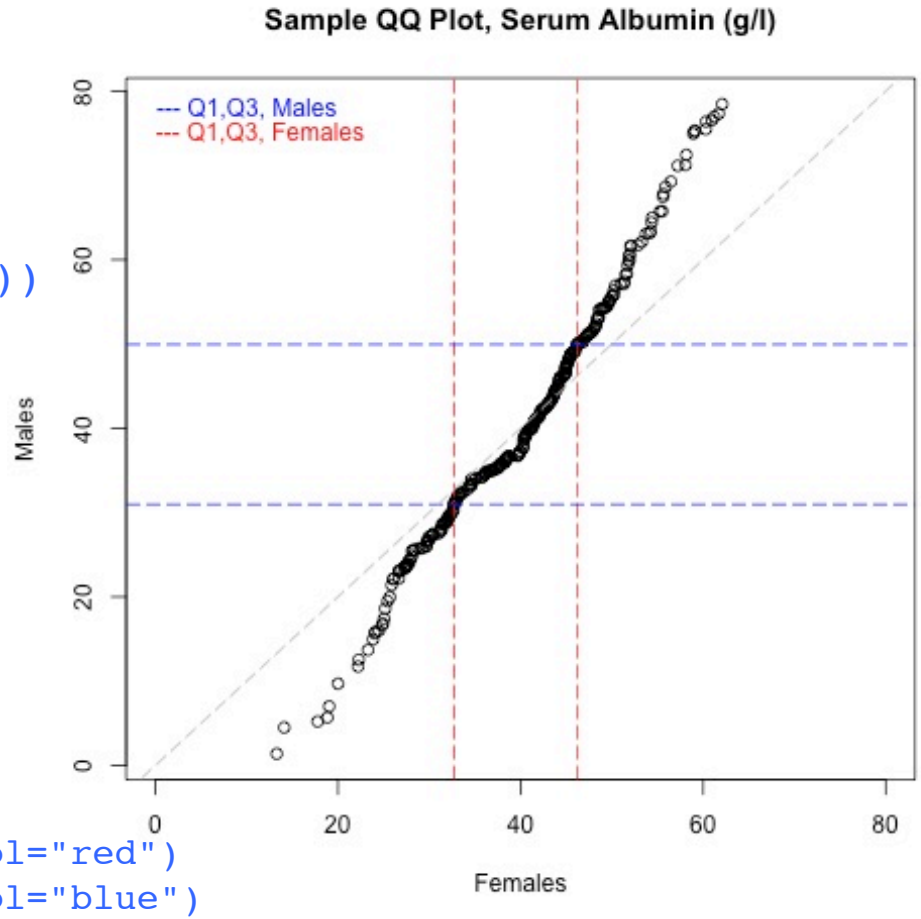
43.43 g/l of Serum Albumin (quantile corresponding to 60th percentile)

Quantile – Quantile plots can be used to:

- Compare two samples
- Compare a sample's distribution with a theoretical distribution (e.g., Normal).

E.g. Sample comparison of Serum Albumin female vs. male

```
> quantile(SA.female,c(.1,.25,.5,.75))
     10%      25%      50%      75%
27.49280 32.72200 41.22050 46.18675

> quantile(SA.male,c(.1,.25,.5,.75))
     10%      25%      50%      75%
23.63270 30.95825 39.95800 49.94775
```




```
qqplot(v1,v2,xlim=c(0,80),ylim=c(0,80))
abline(v=quantile(v1,c(.25,.75)),lty=5,col="red")
abline(h=quantile(v2,c(.25,.75)),lty=5,col="blue")
abline(0,1, lty=5, col="grey")
text(10,78,"--- Q1,Q3, Males",col="blue")
text(10,75,"    --- Q1,Q3, Females",col="red")
```

Exercise

From the `cystfibr` data set,

1. Calculate measures of location and spread for the maximal expiratory pressure (`pemax`) variable separately for males and females.

2. Make a comparative graph with box plots (`range=1.5`) side by side corresponding to the `pemax` by gender. What can you say from the two samples?

3. Make a histogram of the `pemax` variable and draw lines indicating the quartiles Q1 and Q3. Hint: use the `abline()` function to draw the lines, e.g., for Q1:

```
abline(v=quantile(variable,.25),  lty=1.5,  col="red")
```

4. Reproduce the histogram and cumulative histograms for the Hypothetical Serum Albumin data. Use the R code in the slides for help.