Introduction to Fundemental Statitsics in R

Overall Goal

- Create a Table 1 with corresponding plots

Learning objectives:

- Descriptive statistics (mean, median, variance)
- t-tests, Wilcoxon rank sum
- ANOVA, Kruskal Wallis
- Chi-square tests
- Correlations, heat map
- Plots: density plot, box plot, histogram, heat map

Commenting

- Use # signs to comment.
- Anything to the right of a # is ignored by R.
- Helpful to comment code.

Libraries

- We will use the boot package.

```
# Use the commented command below if you do not have boot installed
# install.packages("boot")
library(boot)

# Use the commented command below if you do not have table1 installed
#install.packages("table1")
library(table1)

#install.packages("corrplot")
library(corrplot)
```

Dataset

- Load datasets using read.table, read.csv, fread (big datasets), etc
- Make sure you are in the right directory. Either use the path to load the data or pick the working directory where the data is stored (Session-Set Working directory -> or Choose directory or Misc-> change working directory)
- Example:

```
# d1<-read.table("/Users/path/datasetname.txt", sep="\t",header=T, na.strings="")
```

- Can store and load data as rdata but may be cumbersome, but allows for more complex data formats

Melanoma dataset

- ?melanoma
- Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993) Statistical Models Based on Counting Processes. Springer-Verlag.

The data consist of measurements made on patients with malignant melanoma.

- Each patient had their tumour removed by surgery at the Department of Plastic Surgery, University Hospital of Odense, Denmark during the period 1962 to 1977.
- Surgery consisted of complete removal of the tumour together with about 2.5cm of surrounding skin.
- Among the measurements taken were the thickness of the tumour and whether it was ulcerated or not. These are thought to be important prognostic variables in that patients with a thick and/or ulcerated tumour have an increased chance of death from melanoma.
- Patients were followed until the end of 1977.

This data frame contains the following columns:

- time: Survival time in days since the operation, possibly censored.
- status: The patients status at the end of the study. 1 indicates that they had died from melanoma, 2 indicates that they were still alive and 3 indicates that they had died from causes unrelated to their melanoma.
- sex: The patients sex; 1=male, 0=female.
- age: Age in years at the time of the operation.
- year: Year of operation.
- thickness: Tumour thickness in mm.
- ulcer: Indicator of ulceration; 1=present, 0=absent.

What information would we need?

- Number/percent of subjects with each status
- Mean/var of age and thickness by status or median/range
- Maybe test is mean age differs by status (t-test/ANOVA)
- Number/percent for sex and ulcer by status
- Maybe test is status differs by sex (chi-square test)
- Consider if age is associated with sex (correlation)

First, examine the dataset.

```r
# create a melanoma2 dataset based on melanoma (alternative: melanoma2 = melanoma)
melanoma2 <- melanoma

# dimensions of melanoma2 matrix
dim(melanoma2)
```

```
## [1] 205   7
```

```r
# Check if there is missing data
# melanoma2<-na.omit(melanoma2) #creates data set without missing values
dim(na.omit(melanoma2))
```

```
## [1] 205   7
```

```
# view first 3 rows
melanoma2[1:3,]
```

```
##   time status sex age year thickness ulcer
## 1   10      3   1  76 1972      6.76     1
## 2   30      3   1  56 1968      0.65     0
## 3   35      2   1  41 1977      1.34     0
```

```
# view first 3 rows and first 2 columns
melanoma2[1:3,1:2]
```

```
##   time status
## 1   10      3
## 2   30      3
## 3   35      2
```

```
# column names of variables
colnames(melanoma2)
```

```
## [1] "time"      "status"    "sex"       "age"       "year"      "thickness"
## [7] "ulcer"
```

```
# Use $ for a given variable or melanoma2[,"age"]
melanoma2$age
```

```
##   [1] 76 56 41 71 52 28 77 60 49 68 53 64 68 63 14 72 46 72 95 54 89 25 37 43 68
##  [26] 67 86 56 16 42 65 52 58 60 68 75 19 66 56 46 58 74 65 64 27 73 56 63 69 77
##  [51] 80 76 65 61 26 57 45 31 36 46 43 68 57 57 55 58 20 67 44 59 32 83 55 15 58
##  [76] 47 54 55 38 41 56 48 44 70 40 53 65 54 71 49 55 69 83 60 40 77 35 46 34 69
## [101] 60 84 66 56 75 36 52 58 39 68 71 52 55 66 35 44 72 58 54 33 45 62 72 51 77
## [126] 43 65 63 60 50 40 67 69 74 49 47 42 54 72 45 67 48 34 44 31 42 24 58 78 62
## [151] 70 35 61 54 29 64 47 62 32 49 25 49 64 36 58 37 54 61 31 61 60 43 68  4 60
## [176] 50 20 54 29 56 60 46 42 34 56 12 21 46 49 35 42 47 69 52 52 30 22 55 26 19
## [201] 29 40 42 50 41
```

- Determine the number of subjects

```
# How many subjects?
nrow(melanoma2)
```

```
## [1] 205
```

```
# How many subjects with status 1: died from melanoma? Percent?
nrow(melanoma2[melanoma2$status==1,])
```

```
## [1] 57
```

```
nrow(melanoma2[melanoma2$status==1,])/nrow(melanoma2)
```

```
## [1] 0.2780488
```

```
# How many subjects with status 2: alive? Percent?
length(melanoma2$status[melanoma2$status==2])
```

```
## [1] 134
```

```
length(melanoma2$status[melanoma2$status==2])/length(melanoma2$status)
```

```
## [1] 0.6536585
```

```r
# How many subjects with status 3 and did not have missing data?
nrow(melanoma2[melanoma2$status==3 & !is.na(melanoma2$status),])
```

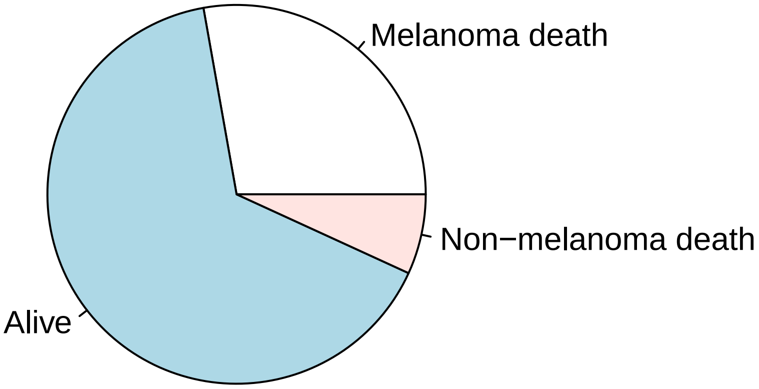
```
## [1] 14
```

```r
#View status as a piechart
x<-c(nrow(melanoma2[melanoma2$status==1,]),nrow(melanoma2[melanoma2$status==2,]),
        nrow(melanoma2[melanoma2$status==3,]))
x
```

```
## [1]  57 134  14
```

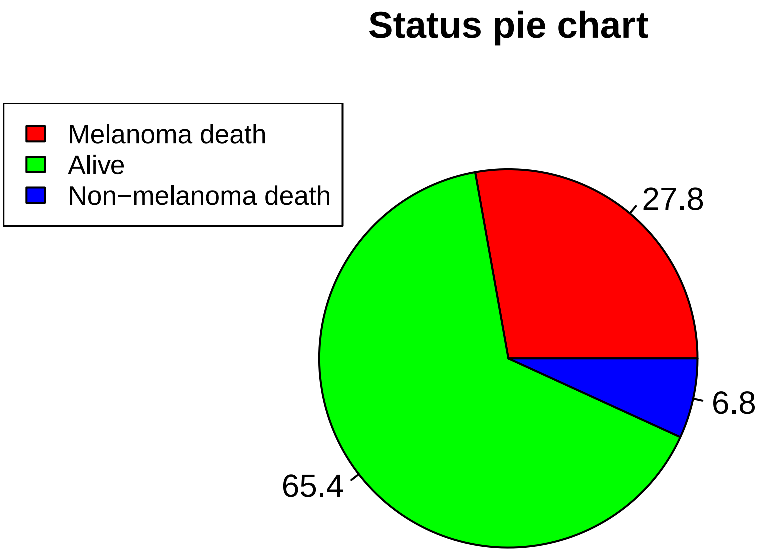
```r
labels=c("Melanoma death","Alive","Non-melanoma death")

#regular pie chart
pie(x,labels)
```



```r
# pie chart with percents
piepercent<- round(100*x/sum(x), 1)

pie(x, labels = piepercent, main = "Status pie chart",col = rainbow(length(x)))
legend("topleft", c("Melanoma death","Alive","Non-melanoma death"), cex = 0.8,
   fill = rainbow(length(x)))
```



Mean, standard deviation, variance, median

- Consider age. What is the average age?

```
# mean for age overall or sum(melanoma2$age)/length(melanoma2$age)
mean(melanoma2$age)
```

```
## [1] 52.46341
```

```
#mean age if there where missing values in dataset
mean(melanoma2$age,na.rm=T)
```

```
## [1] 52.46341
```

```
# round mean age to 1 decimal place
round(mean(melanoma2$age),1)
```

```
## [1] 52.5
```

```
# mean age for status 1: died from melanoma
mean(melanoma2$age[melanoma2$status==1])
```

```
## [1] 55.08772
```

```
# mean age for status 2: alive
mean(melanoma2$age[melanoma2$status==2])
```

```
## [1] 50.00746
```

```
# mean age for status 3: died not from melanoma
mean(melanoma2$age[melanoma2$status==3])
```

```
## [1] 65.28571
```

- What is the variance and standard deviation?

```
# sd for age overall
sd(melanoma2$age)
```

```
## [1] 16.67171
```

```
# sd for age if there where missing values in dataset
sd(melanoma2$age,na.rm=T)
```

```
## [1] 16.67171
```

```
# variance for age overall
sd(melanoma2$age)^2
```

```
## [1] 277.946
```

```
var(melanoma2$age)
```

```
## [1] 277.946
```

```
# variance for age by status
round(c(var(melanoma2$age[melanoma2$status==1]),var(melanoma2$age[melanoma2$status==2]),
        var(melanoma2$age[melanoma2$status==3])),1)
```

```
## [1] 320.7 253.3 118.8
```

- What is the median and range for age?

```
# median for age overall
median(melanoma2$age)
```

```
## [1] 54
```

```
# minimum age
min(melanoma2$age)
```

```
## [1] 4
```

```
#maximum age
max(melanoma2$age)
```

```
## [1] 95
```

```
#range
range(melanoma2$age)
```

```
## [1]  4 95
```

```
# create new variables
AgeStatus1<-melanoma2$age[melanoma2$status==1]
AgeStatus2<-melanoma2$age[melanoma2$status==2]
AgeStatus3<-melanoma2$age[melanoma2$status==3]

#median age by status
c(median(AgeStatus1),median(AgeStatus2),median(AgeStatus3))
```

```
## [1] 56 52 65
```

```
#matrix of ranges
matRange<-matrix(c(range(AgeStatus1),range(AgeStatus2),range(AgeStatus3)),nrow=2,ncol=3)
colnames(matRange)<-c("Status1","Status2","Status3")
rownames(matRange)<-c("min","max")
matRange
```

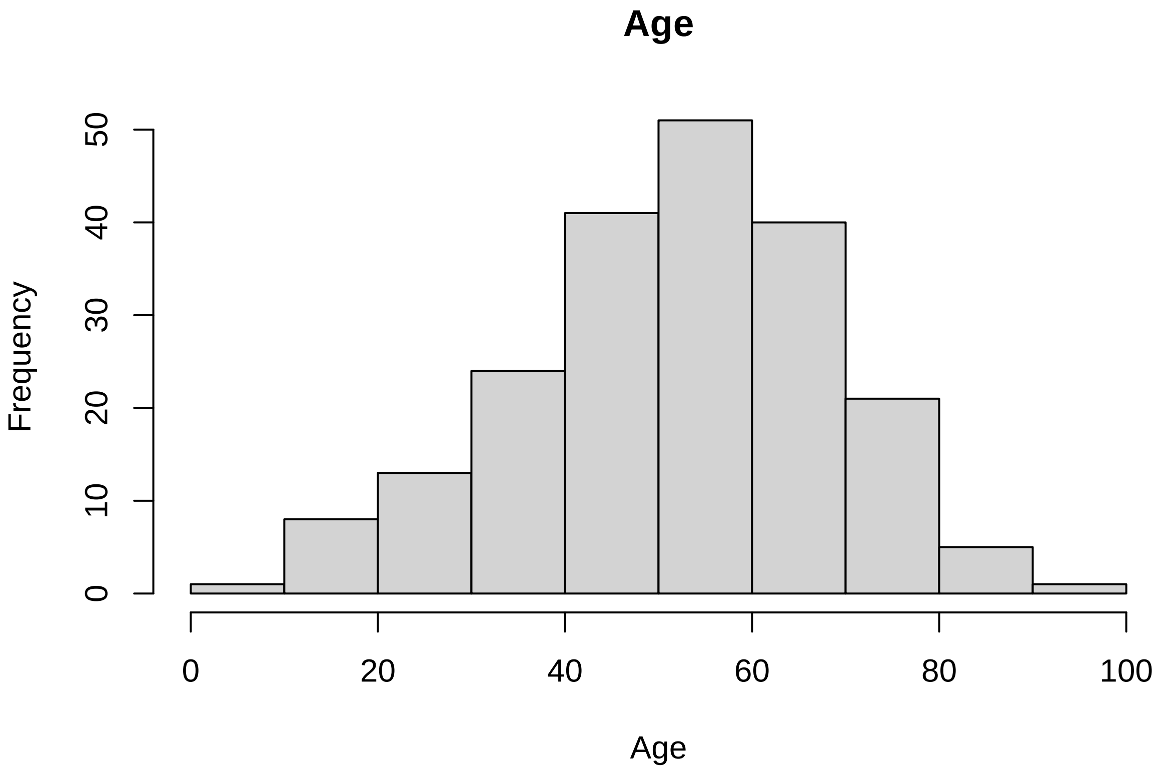
```
##     Status1 Status2 Status3
## min      14       4      49
## max      95      84      86
```

```
range(AgeStatus1)
```

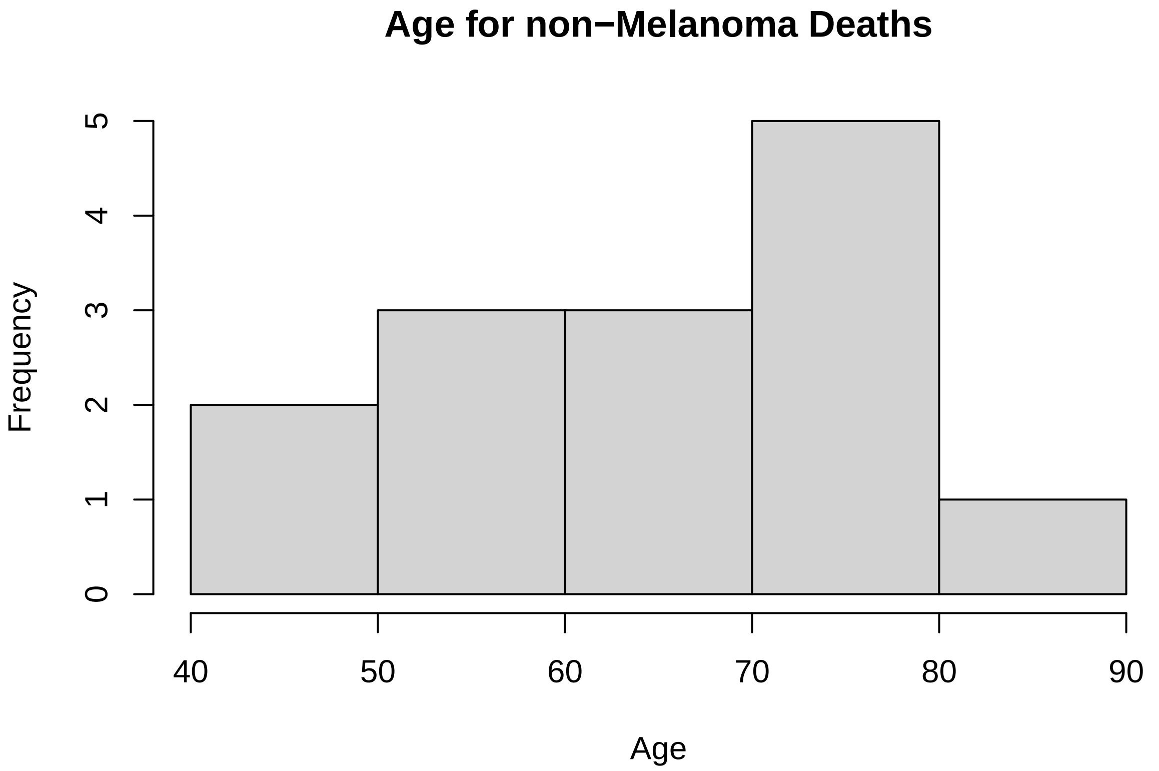
```
## [1] 14 95
```

- Should we be using medians or means? Check normality (Kolmogorov–Smirnov or Shapiro–Wilk test)

```
hist(melanoma2$age,main="Age",xlab="Age")
```

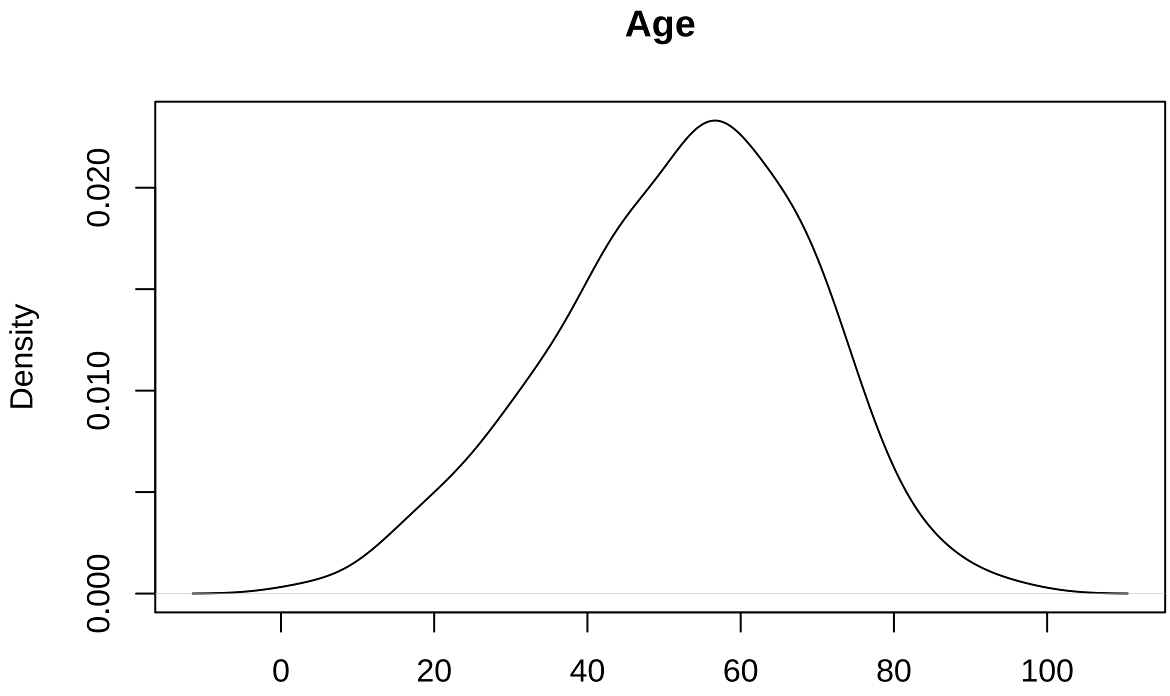


```
hist(melanoma2$age[melanoma2$status==3],main="Age for non-Melanoma Deaths",xlab="Age")
```

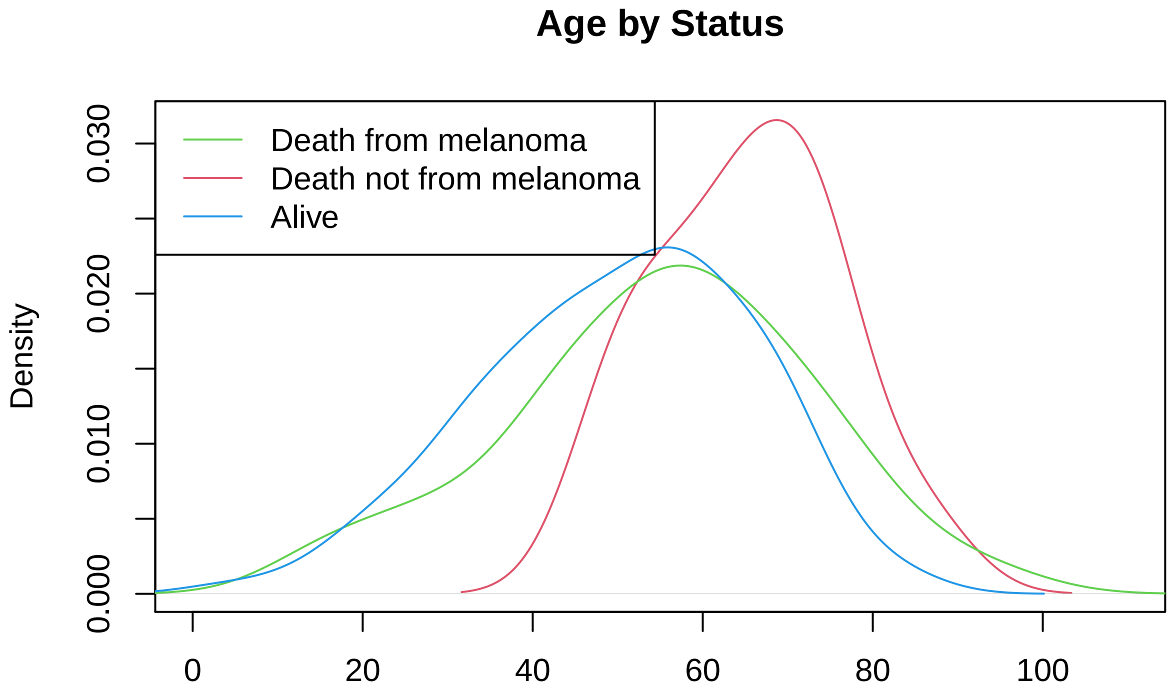


- Density plots (status samples size= 57,134,14)

```
#plot density for age overall
plot(density(melanoma2$age),main="Age",xlab="")
```

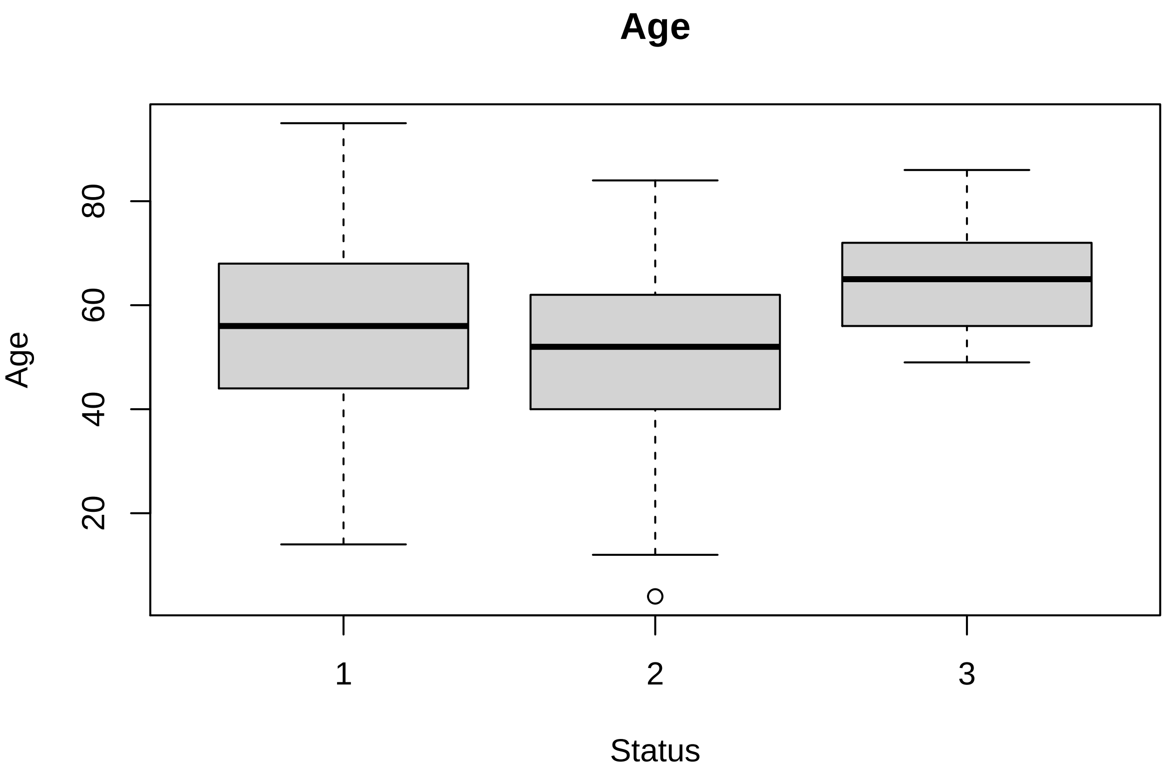


```
#plot density for age by status
plot(density(melanoma2$age[melanoma2$status==3]),main="Age by Status",xlab="",xlim=c(0,110),col=2)
lines(density(melanoma2$age[melanoma2$status==1]),col=3)
lines(density(melanoma2$age[melanoma2$status==2]),col=4)
legend("topleft",c("Death from melanoma","Death not from melanoma","Alive"),col=c(3,2,4),lwd=1)
```



- Boxplot

```
boxplot(age~status,data=melanoma2, main="Age",
   xlab="Status", ylab="Age")
```



t-tests and Wilcoxon rank sum test

- Does the average age differ by status 1 and 2?
- Which test to use? t-test assumes normality. t-test with equal variance or unequal?

```
# t test
t.test(melanoma2$age[melanoma2$status==1],melanoma2$age[melanoma2$status==2],
       alternative = "two.sided",paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  melanoma2$age[melanoma2$status == 1] and melanoma2$age[melanoma2$status == 2]
## t = 1.853, df = 95.424, p-value = 0.06698
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3623352 10.5228485
## sample estimates:
## mean of x mean of y
##  55.08772  50.00746
```

- p-value=0.07. Average age does not significantly differ for subjects who died by melanoma versus subjects who were still alive.

- Wilcoxon rank sum test may be approriate if normality assumption is not met or samlple size is small. Used to test if the median differs in the 2 groups.

```
#Wilcoxon rank sum exact test (signed rank for paired =True)
wilcox.test(melanoma2$age[melanoma2$status==1],melanoma2$age[melanoma2$status==2],
            alternative = "two.sided",paired = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  melanoma2$age[melanoma2$status == 1] and melanoma2$age[melanoma2$status == 2]
## W = 4458.5, p-value = 0.06749
## alternative hypothesis: true location shift is not equal to 0
```

ANOVA and Kruskal-Wallis

- Does the average age differ by status?
- ANOVA assumes normality and vairance equal in all 3 groups.
- The Kruskal-Wallis test is a non parametric test used to test the null hypothesis which states that 'k' number of samples has been drawn from the same population or the identical population with the same or identical median.

```
#ANOVA
oneway.test(age ~ status,data = melanoma2,var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  age and status
## F = 6.6498, num df = 2, denom df = 202, p-value = 0.001596
```

```
#Kruskall-Wallis
kruskal.test(age ~ status,data = melanoma2)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  age by status
## Kruskal-Wallis chi-squared = 13.028, df = 2, p-value = 0.001483
```

- p-value = 0.00159. The average age differs by status for at least 2 groups.

Chi-square tests

- Is sex associated with status? Yes, p-value = 0.03.

```
table(melanoma2$sex, melanoma2$status)
```

```
##
##      1  2  3
##   0 28 91  7
##   1 29 43  7
```

```
chisq.test(melanoma2$sex, melanoma2$status)
```

```
##
##  Pearson's Chi-squared test
##
## data:  melanoma2$sex and melanoma2$status
## X-squared = 6.793, df = 2, p-value = 0.03349
```

Correlations

- Is age correlated with sex?

```
cor(melanoma2$sex,melanoma2$age)
```

```
## [1] 0.06833741
```

```
# Test correlation (pearson, could use method = "spearman")
cor.test(melanoma2$sex,melanoma2$age,alternative = "two.sided",method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  melanoma2$sex and melanoma2$age
## t = 0.97594, df = 203, p-value = 0.3303
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.06934701  0.20346704
## sample estimates:
##        cor
## 0.06833741
```

- Age is note significantly correlated with sex (p-value=0.33).

- Plots for correlation. Need one of the traits to be continuous.

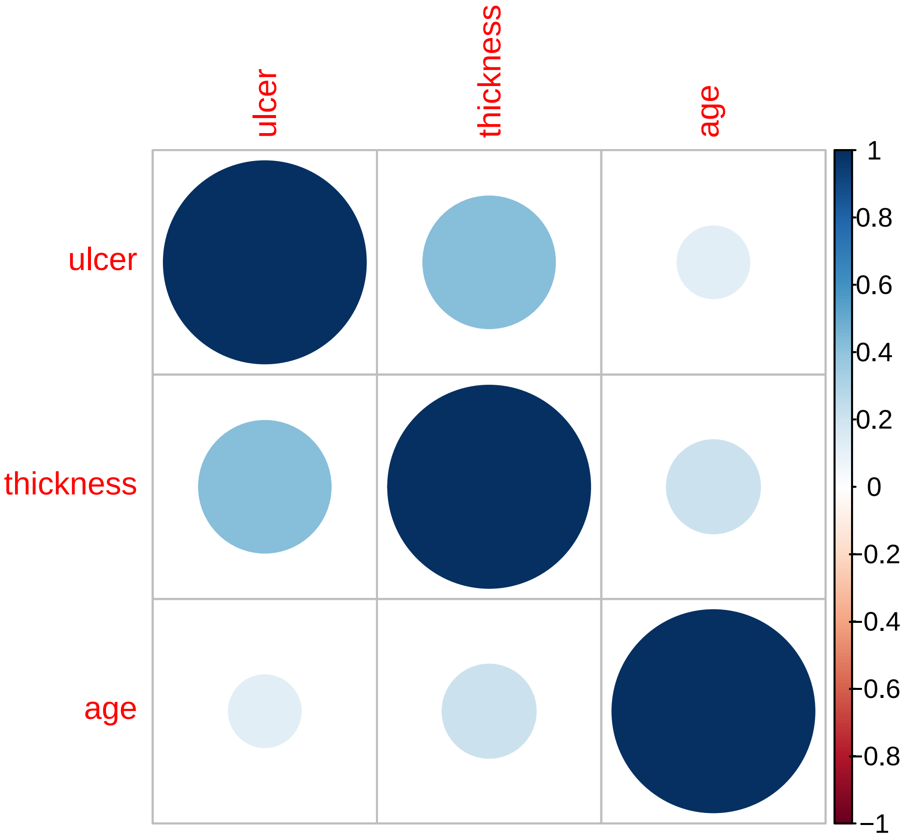
```
#https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

# Create new matrix with only age, thickness, ulcer
melanoma3<-melanoma2[,c("age","thickness","ulcer")]

# Check
melanoma3[1:5,]
```

```
##   age thickness ulcer
## 1  76      6.76     1
## 2  56      0.65     0
## 3  41      1.34     0
## 4  71      2.90     0
## 5  52     12.08     1
```

```
# Plot correlation
M = cor(melanoma3)
corrplot(M, order = 'AOE')
```



Existing packages

```r
# https://cran.r-project.org/web/packages/table1/vignettes/table1-examples.html

# change status variable to text
melanoma2$status <-
   factor(melanoma2$status,
          levels=c(2,1,3),
          labels=c("Alive", # Reference
                   "Melanoma death",
                   "Non-melanoma death"))
melanoma2[1:5,]
```

```
##   time             status sex age year thickness ulcer
## 1   10 Non-melanoma death   1  76 1972      6.76     1
## 2   30 Non-melanoma death   1  56 1968      0.65     0
## 3   35              Alive   1  41 1977      1.34     0
## 4   99 Non-melanoma death   0  71 1968      2.90     0
## 5  185     Melanoma death   1  52 1965     12.08     1
```

```r
table1(~ factor(sex) + age + factor(ulcer) + thickness | status, data=melanoma2)
```

	Alive	Melanoma death	Non-melanoma death	Overall
	(N=134)	(N=57)	(N=14)	(N=205)
factor(sex)				
0	91 (67.9%)	28 (49.1%)	7 (50.0%)	126 (61.5%)
1	43 (32.1%)	29 (50.9%)	7 (50.0%)	79 (38.5%)
age				
Mean (SD)	50.0 (15.9)	55.1 (17.9)	65.3 (10.9)	52.5 (16.7)
Median [Min, Max]	52.0 [4.00, 84.0]	56.0 [14.0, 95.0]	65.0 [49.0, 86.0]	54.0 [4.00, 95.0]
factor(ulcer)				
0	92 (68.7%)	16 (28.1%)	7 (50.0%)	115 (56.1%)
1	42 (31.3%)	41 (71.9%)	7 (50.0%)	90 (43.9%)
thickness				
Mean (SD)	2.24 (2.33)	4.31 (3.57)	3.72 (3.63)	2.92 (2.96)
Median [Min, Max]	1.36 [0.100, 12.9]	3.54 [0.320, 17.4]	2.26 [0.160, 12.6]	1.94 [0.100, 17.4]

R resources

- Harvard Catalyst or other free Harvard courses https://online-learning.harvard.edu/subject/r
- There are online courses through coursera https://www.coursera.org/learn/r-programming
- Software carpentry offers really fun 2 day workshops. You can check when there is one in Boston and make sure to sign up right away because they fill up quickly https://software-carpentry.org/workshops/