STATS 3001 / STATS 4104 / STATS 7054
Statistical Modelling III
Exam

E

```
# clear all variables, functions, etc
# clean up memory
rm(list=ls())
# clean up memory
gc()
```

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 471281 25.2    1017511 54.4   658011 35.2
## Vcells 877206  6.7    8388608 64.0  1769872 13.6
```

loading data

```
# load mtcars data
data(mtcars)
# show the data in the dataframe
mtcars
```

```
##                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4          21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag      21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710         22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive     21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant            18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360         14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D          24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230           22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280           19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C          17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE         16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL         17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC        15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4  8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial  14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128           32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic        30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
```

```
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

readr::read_csv("mtcars.csv") getwd() save(mtcars, file = "mtcars.RData") load("mtcars.RData")

Apply family

```
m <- matrix(1:12, 3, 4)
m
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    4    7   10
## [2,]    2    5    8   11
## [3,]    3    6    9   12
```

```
sum(m)
```

```
## [1] 78
```

```
# apply function to each row
apply(m, 1, sum)
```

```
## [1] 22 26 30
```

```
# apply function to each column
apply(m, 2, sum)
```

```
## [1]  6 15 24 33
```

```
# apply function to each row
rowSums(m)
```

```
## [1] 22 26 30
```

```
apply(m, 1, mean)
```

```
## [1] 5.5 6.5 7.5
```

```r
rowMeans(m)
```

```
## [1] 5.5 6.5 7.5
```

```r
apply(m, 2, sqrt)
```

```
##          [,1]     [,2]     [,3]     [,4]
## [1,] 1.000000 2.000000 2.645751 3.162278
## [2,] 1.414214 2.236068 2.828427 3.316625
## [3,] 1.732051 2.449490 3.000000 3.464102
```

```r
# apply is for matrices
apply(mtcars, 2, mean)
```

```
##        mpg        cyl       disp         hp       drat         wt       qsec
##  20.090625   6.187500 230.721875 146.687500   3.596563   3.217250  17.848750
##         vs         am       gear       carb
##   0.437500   0.406250   3.687500   2.812500
```

```r
# sapply is for vectors
sapply(mtcars, mean)
```

```
##        mpg        cyl       disp         hp       drat         wt       qsec
##  20.090625   6.187500 230.721875 146.687500   3.596563   3.217250  17.848750
##         vs         am       gear       carb
##   0.437500   0.406250   3.687500   2.812500
```

```r
# lapply is for lists
lapply(mtcars, mean)
```

```
## $mpg
## [1] 20.09062
##
## $cyl
## [1] 6.1875
##
## $disp
## [1] 230.7219
##
## $hp
## [1] 146.6875
##
## $drat
## [1] 3.596563
##
## $wt
## [1] 3.21725
##
## $qsec
## [1] 17.84875
```

```
##
## $vs
## [1] 0.4375
##
## $am
## [1] 0.40625
##
## $gear
## [1] 3.6875
##
## $carb
## [1] 2.8125
```

Descriptive statistics

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
summary(mtcars[,c('mpg', 'wt')])
```

```
##       mpg             wt
##  Min.   :10.40   Min.   :1.513
##  1st Qu.:15.43   1st Qu.:2.581
##  Median :19.20   Median :3.325
##  Mean   :20.09   Mean   :3.217
##  3rd Qu.:22.80   3rd Qu.:3.610
##  Max.   :33.90   Max.   :5.424
```

```r
mean(mtcars$mpg)
```

```
## [1] 20.09062
```

```r
sd(mtcars$mpg)
```

```
## [1] 6.026948
```

```r
moments::skewness(mtcars$mpg)
```

```
## [1] 0.6404399
```

```r
moments::kurtosis(mtcars$mpg)
```

```
## [1] 2.799467
```

```r
quantile(mtcars$mpg)
```

```
##     0%    25%    50%    75%   100%
## 10.400 15.425 19.200 22.800 33.900
```

```r
table(mtcars$cyl)
```

```
##
##  4  6  8
## 11  7 14
```

```r
temp <- table(mtcars$cyl)
names(temp)
```

```
## [1] "4" "6" "8"
```

```r
unique(mtcars$cyl)
```

```
## [1] 6 4 8
```

```r
table(mtcars$cyl, mtcars$gear)
```

```
##
##      3  4  5
##   4  1  8  2
##   6  2  4  1
##   8 12  0  2
```

Data visualisation

```r
df <- dslabs::us_contagious_diseases

head(df)
```

```
##       disease   state year weeks_reporting count population
## 1 Hepatitis A Alabama 1966              50   321    3345787
## 2 Hepatitis A Alabama 1967              49   291    3364130
## 3 Hepatitis A Alabama 1968              52   314    3386068
## 4 Hepatitis A Alabama 1969              49   380    3412450
## 5 Hepatitis A Alabama 1970              51   413    3444165
## 6 Hepatitis A Alabama 1971              51   378    3481798
```

```r
names(df)
```

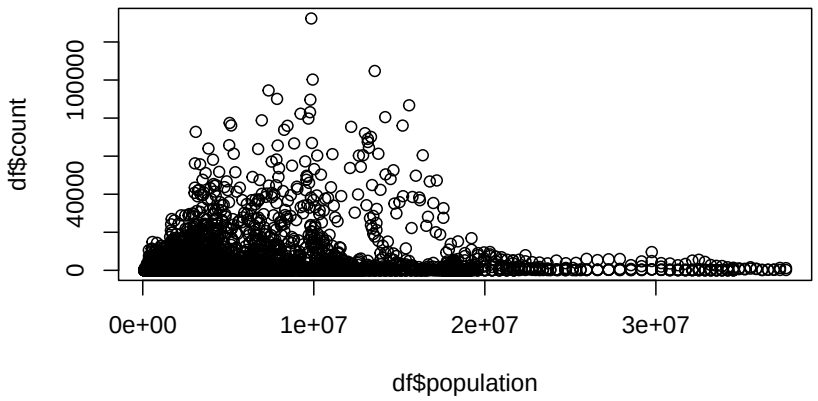
```
## [1] "disease"         "state"           "year"            "weeks_reporting"
## [5] "count"           "population"
```

```r
dim(df)
```

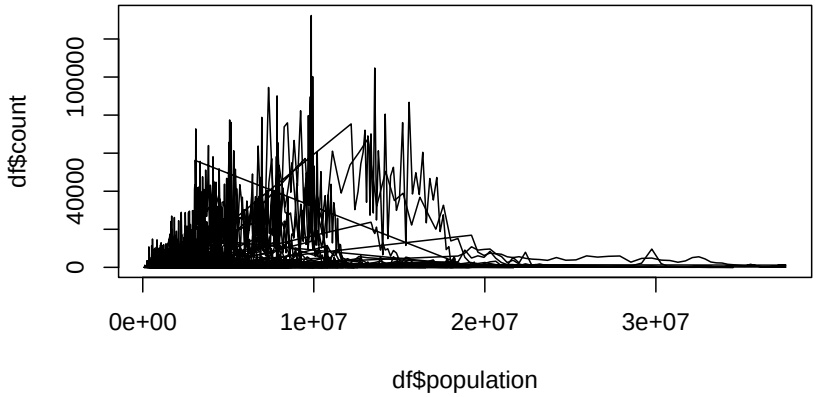
```
## [1] 16065     6
```

?dslabs::us_contagious_diseases

```r
plot(df$population, df$count)
```

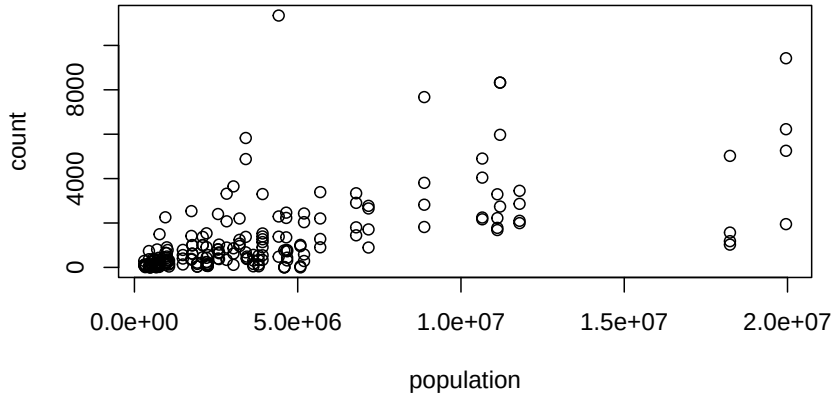



```r
plot(df$population, df$count, type = "l")
```

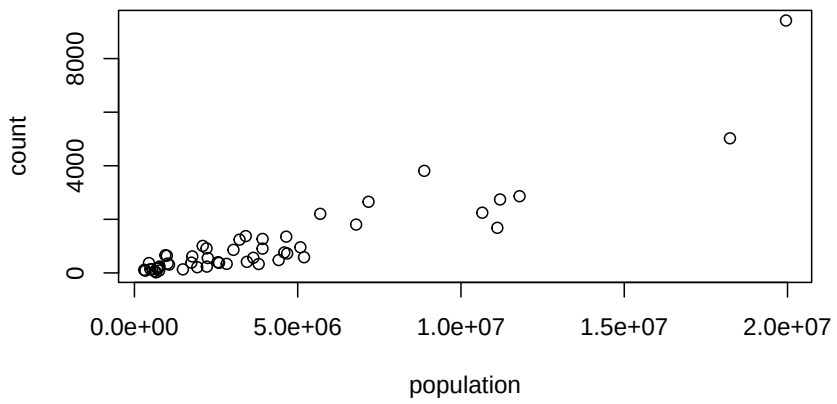



```
x <- df$year == 1970
dfx <- df[x, ]
```

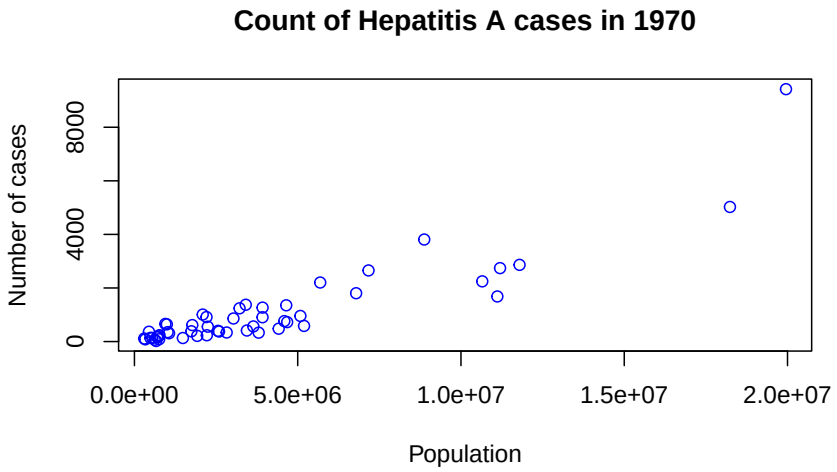
```
plot(df[df$year == 1970, c("population", "count")])
```



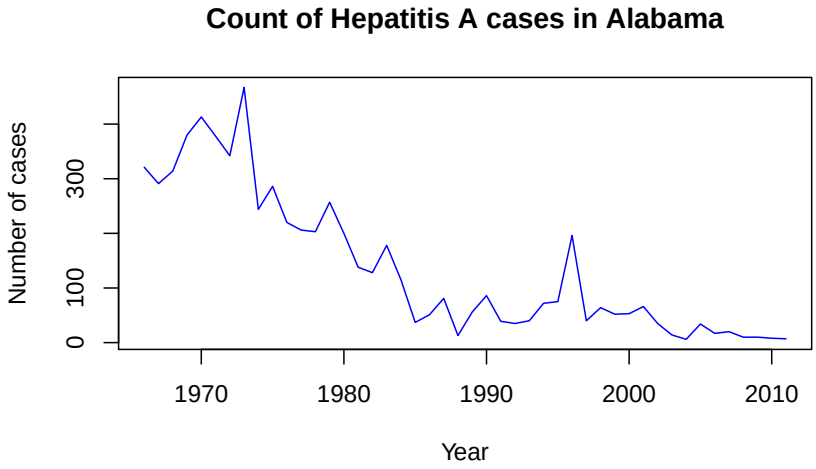

```
plot(df[df$year == 1970 & df$disease == "Hepatitis A", c("population", "count")])
```



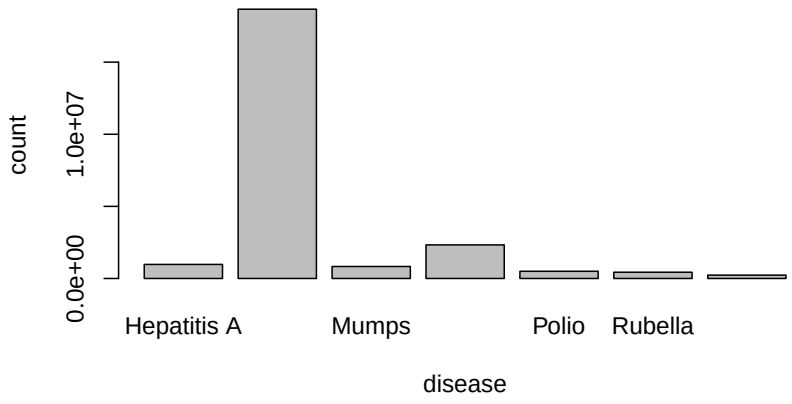

```
plot(df[df$year == 1970 & df$disease == "Hepatitis A", c("population", "count")],
     main = "Count of Hepatitis A cases in 1970",
     xlab = "Population", ylab = "Number of cases", col='blue')
```



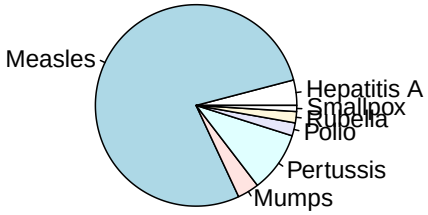

```
# plot the count of Hepatitis A cases in Alabama
df_small <- df[df$disease == "Hepatitis A" & df$state == "Alabama", c("year", "count")]
plot(df_small, type = 'l', main = "Count of Hepatitis A cases in Alabama",
     xlab = "Year", ylab = "Number of cases", col='blue')
```




```
# aggregate the data by disease
disease_cases <- aggregate(df$count, by = list(df$disease), FUN = sum)
# rename the columns, so that they are more meaningful
names(disease_cases) <- c("disease", "count")
# plot the count of cases by disease
barplot(count ~ disease, data = disease_cases)
```

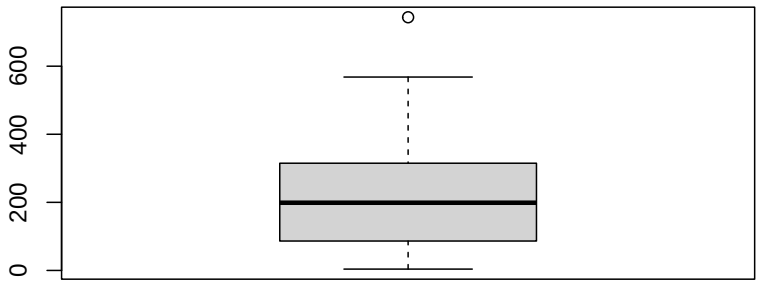



```
pie(disease_cases$count, labels = disease_cases$disease)
```

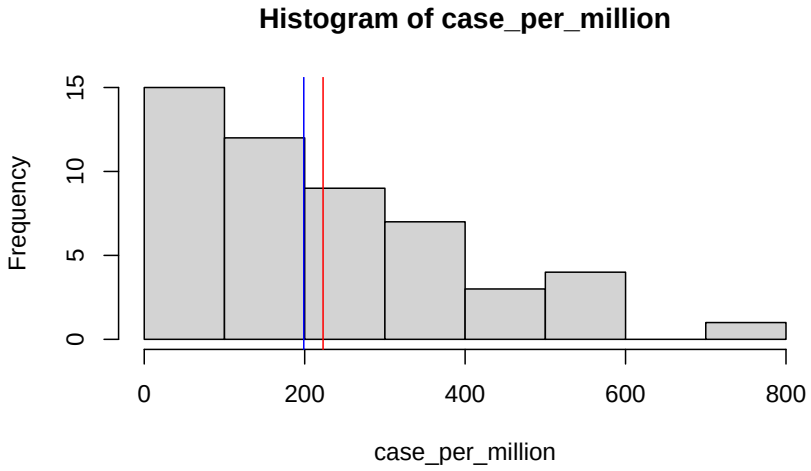



```
# normalise by population
df$case_per_million <- df$count / df$population * 10^6
# select the data for measles in 1970
case_per_million <- df[df$disease == "Measles" & df$year == 1970, c("case_per_million")]

boxplot(case_per_million)
```




```
hist(case_per_million)
# add a vertical line for the mean
abline(v = mean(case_per_million), col = "red")
# add a vertical line for the median
abline(v = median(case_per_million), col = "blue")
```




```
boxplot(case_per_million)
abline(h = mean(case_per_million), col = "red")
```