Homework 10

Maya Mileva

due date: Dec 12th, 2019

I did this homework by myself, with help from the book and the professor.

Exercises

2. Download and library the nlme package and use data ("Blackmore") to activate the Blackmore data set. Inspect the data and create a box plot showing the exercise level at different ages. Run a repeated measures ANOVA to compare exercise levels at ages 8, 10, and 12 using aov(). You can use a command like, myData <-Blackmore[Blackmore$age <=12,], to subset the data. Keeping in mind that the data will need to be balanced before you can conduct this analysis, try running a command like this, table(myData$subject, myData$age)), as the starting point for cleaning up the data set.

The Blackmore data frame has 945 rows and 4 columns. Blackmore and Davis's data on exercise histories of 138 teenaged girls hospitalized for eating disorders and 98 control subjects.

This data frame contains the following columns:

subject

a factor with subject id codes. There are several observations for each subject, but because the girls were hospitalized at different ages, the number of cases and the age at the last case vary.

age

subject's age in years at the time of observation; all but the last observation for each subject were collected retrospectively at intervals of two years, starting at age 8.

exercise

the amount of exercise in which the subject engaged, expressed as estimated hours per week.

group

a factor with levels: control, Control subjects; patient, Eating-disordered patients.

```
kable(head(Blackmore),align = 'c')
```

subject	age	exercise	group
100	8.00	2.71	patient
100	10.00	1.94	patient
100	12.00	2.36	patient
100	14.00	1.54	patient
100	15.92	8.63	patient
101	8.00	0.14	patient

```
## Data exploration
str(Blackmore)
```

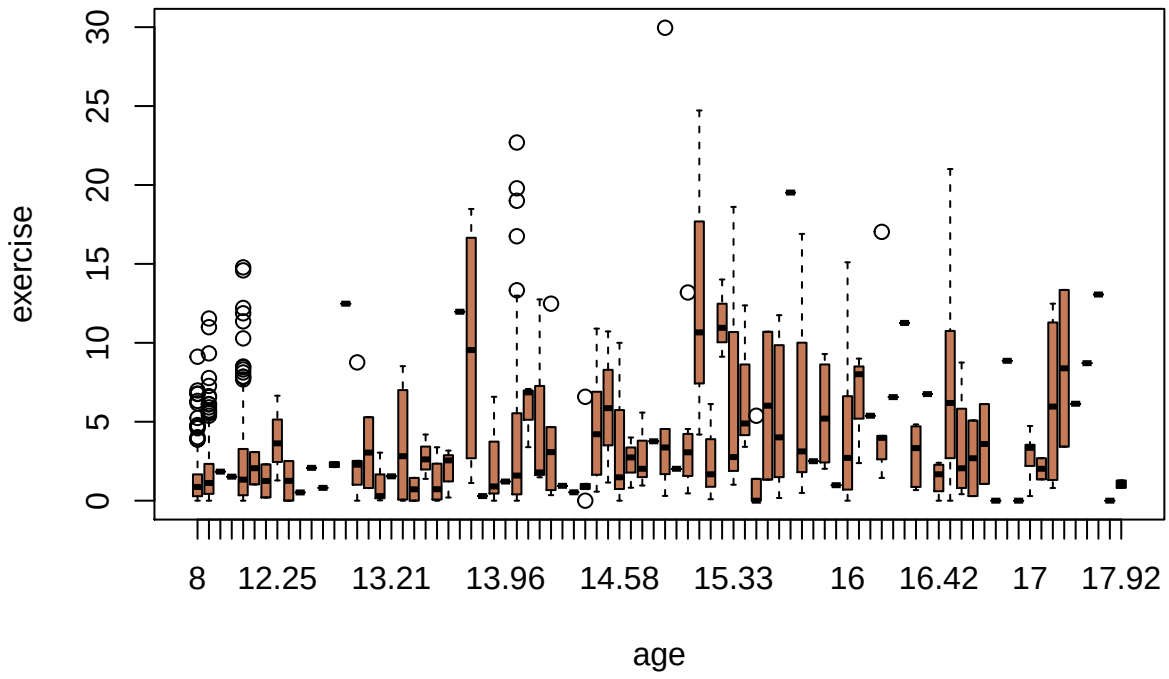
```
## 'data.frame':    945 obs. of  4 variables:
```

```
##  $ subject : Factor w/ 231 levels "100","101","102",..: 1 1 1 1 1 2 2 2 2 2 ...
##  $ age     : num  8 10 12 14 15.9 ...
##  $ exercise: num  2.71 1.94 2.36 1.54 8.63 0.14 0.14 0 0 5.08 ...
##  $ group   : Factor w/ 2 levels "control","patient": 2 2 2 2 2 2 2 2 2 2 ...
```

```
summary(Blackmore)
```

subject	age	exercise	group
100 : 5	Min. : 8.00	Min. : 0.000	control:359
101 : 5	1st Qu.:10.00	1st Qu.: 0.400	patient:586
105 : 5	Median :12.00	Median : 1.330	NA
106 : 5	Mean :11.44	Mean : 2.531	NA
107 : 5	3rd Qu.:14.00	3rd Qu.: 3.040	NA
108 : 5	Max. :17.92	Max. :29.960	NA
(Other):915	NA	NA	NA

```
## Box plot showing the exercise level at different ages
boxplot(exercise~age, data = Blackmore, col = "#C57B57")
```




```
# Create dataframe for ages 8,10,12
myData <- Blackmore[Blackmore$age <= 12,]
myData <- myData[myData$age!=11.58,] # A couple of unusual observations
myData <- myData[myData$age!=11.83,] # A couple of unusual observations
```

```r
table(myData$age) # Check results
```

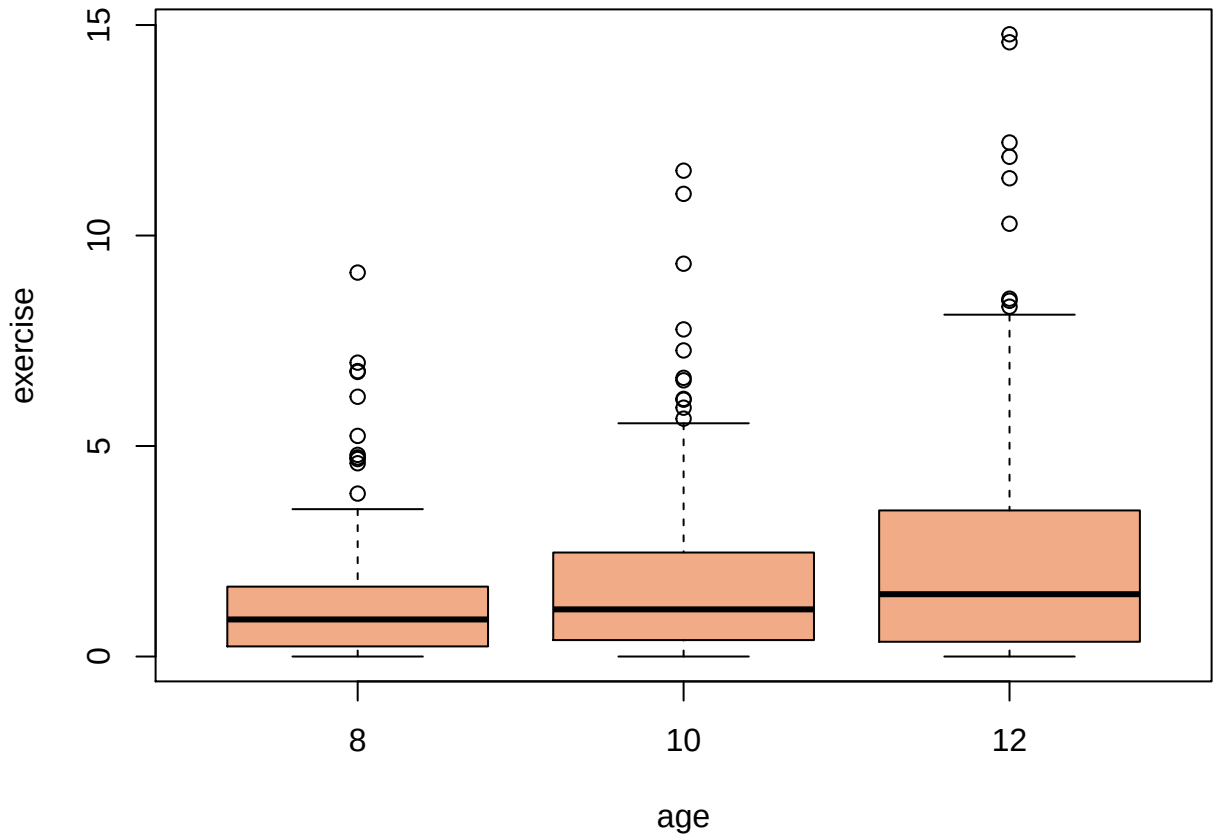
8	10	12
231	229	176

```r
# Balance the data set
myData$ageFact <- as.factor(myData$age) # Make this a factor
list <- rowSums(table(myData$subject, myData$ageFact)) == 3 # Only allow subject IDs with 3 obs
list <- list[list==TRUE] # Squeeze the list down to just the 3s
list <- as.numeric(names(list)) # Eliminate the non-numeric IDs
myData <- myData[myData$subject %in% list, ] # Choose everything on the list
myData$subject <- as.factor(as.numeric(myData$subject)) # Repair the factor levels
# table(myData$ageFact,myData$subject) # Check results
dim(myData)
```

```
## [1] 498   5
```

Repeated-measures ANOVA allows us to compare cases across two or more points in time , rather than just one pair of points. After the data is balanced we can run aov() that will examine the effect of the age(as factor) on exercise level.

```r
par(mar = c(4,4,1,1))
boxplot(exercise ~age, data = myData, col = "#F1AB86")
```



The exercise level increases with the increase of the age from 8 to 10.

To assess whether differences exist in weights across all time groups we need repeated measures ANOVA:

```
summary(aov(exercise ~ ageFact + Error(subject), data = myData))
```

```
##
## Error: subject
##            Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 165   1892   11.47
##
## Error: Within
##            Df Sum Sq Mean Sq F value   Pr(>F)
## ageFact     2  102.7   51.35   27.04 1.33e-11 ***
## Residuals 330  626.6    1.90
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the model formula "exercise ~ age" we are testing the hypothesis that exercise does not vary with the age change. "Error: subject" specifies the individual differences among subjects as error variance that we want to leave out.

If we add up all the degrees of freedom, the total is 497. There are 498 observation in the data set, and after calculating the grand mean there are df = 497 remaining. The sum of the squares is also identical to the one in the data set. What ANOVA table shows is that the variance and the degree of freedom have been partitioned into various components, in pursuit of the goal of separating out the individual differences' variance from the calculation of the F-test.

In the first section of the output, the "Residuals" refers to the variance attributable to individual differences among subjects. The df= 165 signifies that we have 166 subjects in the data set. The sum of the squares 1892 represents variation in the exercise variable that is directly attributable to individual differences among subjects. With repeated measures ANOVA we can separate out this proportion of variation in the dependent variable so that it does not appear in the denominator of the F-test. In the second section the effect of age is expressed as an F-ratio, F (102.7, 626.6) = 27.04, p<.001. The df = 2 reflects to the 3 ages where we measured the exercise level for each of the 498 subjects. The df = 330 in the nominator is the remaining error variance that is not attributable to individual differences. This F-ratio test the null hypothesis that the changes in subject exercise level are consistently 0 across all three ages. With low p-value (1.33e-11) <.001, we can reject the null hypothesis.

```
## Calculate eta-square effect
102.7/(102.7+626.6+1892)
```

```
## [1] 0.03917903
```

The proportion of variance in exercise that is accounted for by age is .04.

Alternative to aov() is ezANOVA.

```
ezANOVA(data = myData, dv=.(exercise),
        within = .(ageFact), wid = .(subject), detailed = T)
```

```
## $ANOVA
##        Effect DFn DFd       SSn       SSd         F            p p<.05
## 1 (Intercept)   1 165 1688.7531 1891.8688 147.28519 1.243957e-24     *
```

```
## 2     ageFact   2 330  102.6938  626.5731  27.04309 1.330658e-11     *
##          ges
## 1 0.40139644
## 2 0.03917912
##
## $`Mauchly's Test for Sphericity`
##    Effect         W            p p<.05
## 2 ageFact 0.7490297 5.115871e-11     *
##
## $`Sphericity Corrections`
##    Effect       GGe        p[GG] p[GG]<.05      HFe        p[HF] p[HF]<.05
## 2 ageFact 0.7993795 9.573171e-10         * 0.805973 8.316244e-10         *
```

The F-test matches what we found with aov(), so we can reject the null hypothesis of no change in exercise with the age increase. ges = 0.039 is also matches our finding. "Mauchly's Test for Sphericity" (test for homogeneity of variance of the differences among pairs of time groups) is significant and that means that we have violated the assumption. When Mauchly's test of sphericity is significant, it raises the possibility that the F-test is incorrect, and we may draw the wrong conclusion from the null hypothesis significance test on F. Fortunately, the ezANOVA() command also provides some additional tests in the third section of the output that we can consult when we have violated the test of sphericity. GGe (with the associated probability p[GG]), applies a correction to the degrees of freedom to counteract the possible inflation of the F-ratio. The associated probability value, p[GG], reflects the same F-ratio evaluated against a more stringent set of degrees of freedom. If the associated p-value remains significant (p = 9.573171e-10), we do not need to revise our decision about rejecting the null hypothesis.

5. Given that the AirPassengers data set has a substantial growth trend, use diff() to create a differenced data set. Use plot() to examine and interpret the results of differencing. Use cpt.var() to find the change point in the variability of the differenced time series. Plot the result and describe in your own words what the change point signifies.

The data set represents monthly totals of international airline passengers, 1949 to 1960.

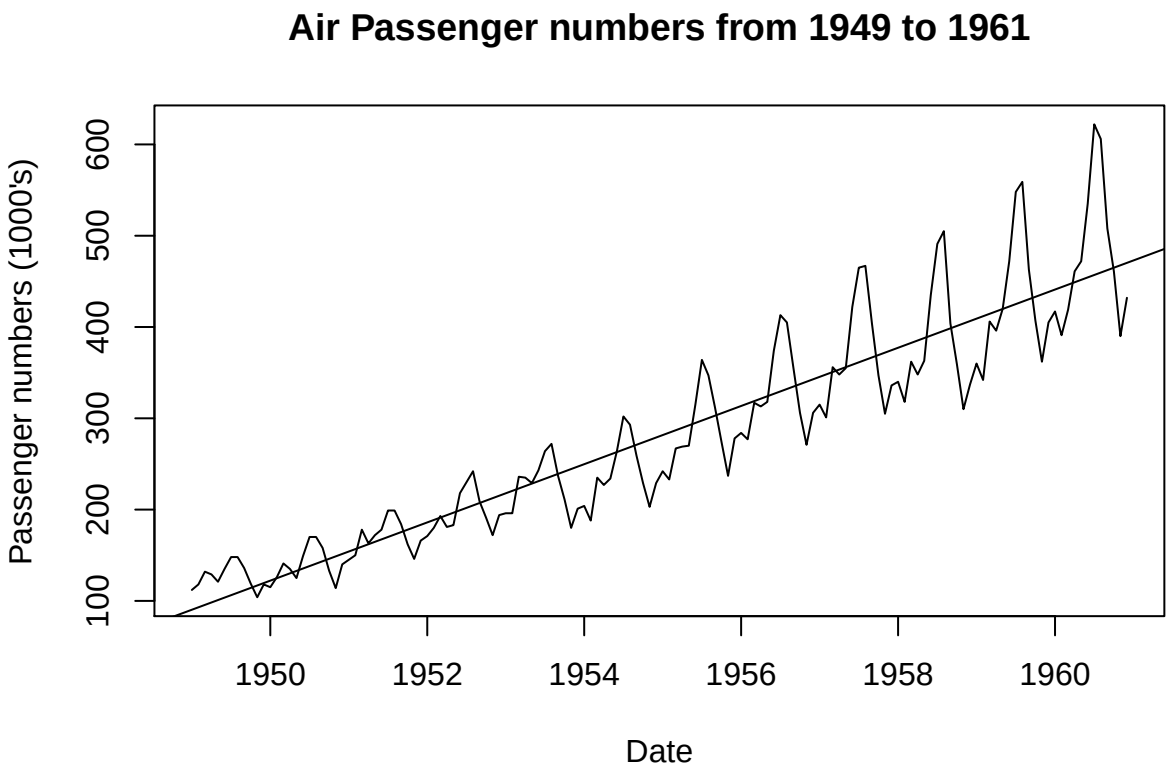
```
# plot(AirPassengers)
class(AirPassengers)
```

```
## [1] "ts"
```

```
sum(is.na(AirPassengers))
```

```
## [1] 0
```

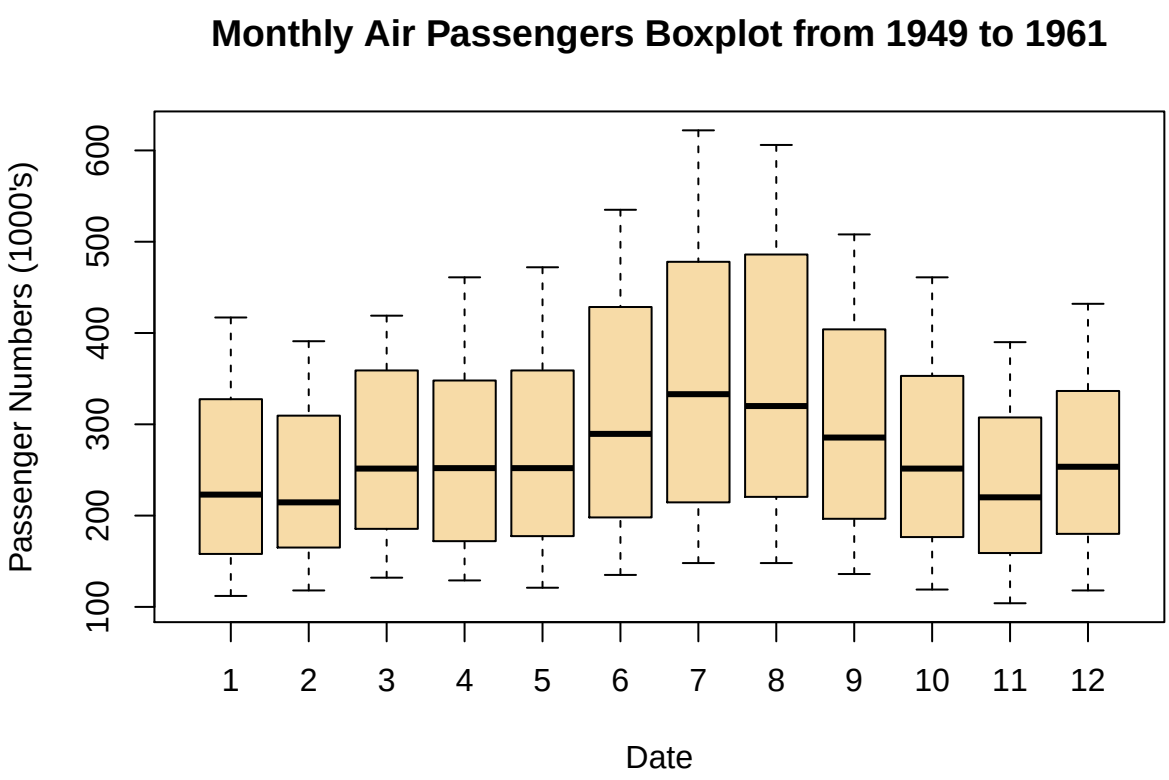
```
plot(AirPassengers, xlab="Date", ylab = "Passenger numbers (1000's)"
     ,main="Air Passenger numbers from 1949 to 1961")# has a substantial growth trend
abline(reg=lm(AirPassengers~time(AirPassengers)))
```




```
cycle(AirPassengers)
```

```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 1949   1   2   3   4   5   6   7   8   9  10  11  12
## 1950   1   2   3   4   5   6   7   8   9  10  11  12
## 1951   1   2   3   4   5   6   7   8   9  10  11  12
## 1952   1   2   3   4   5   6   7   8   9  10  11  12
## 1953   1   2   3   4   5   6   7   8   9  10  11  12
## 1954   1   2   3   4   5   6   7   8   9  10  11  12
## 1955   1   2   3   4   5   6   7   8   9  10  11  12
## 1956   1   2   3   4   5   6   7   8   9  10  11  12
## 1957   1   2   3   4   5   6   7   8   9  10  11  12
## 1958   1   2   3   4   5   6   7   8   9  10  11  12
## 1959   1   2   3   4   5   6   7   8   9  10  11  12
## 1960   1   2   3   4   5   6   7   8   9  10  11  12
```

```
boxplot(AirPassengers~cycle(AirPassengers), col = "#F7DBA7",
        xlab="Date", ylab = "Passenger Numbers (1000's)"
        ,main ="Monthly Air Passengers Boxplot from 1949 to 1961")
```

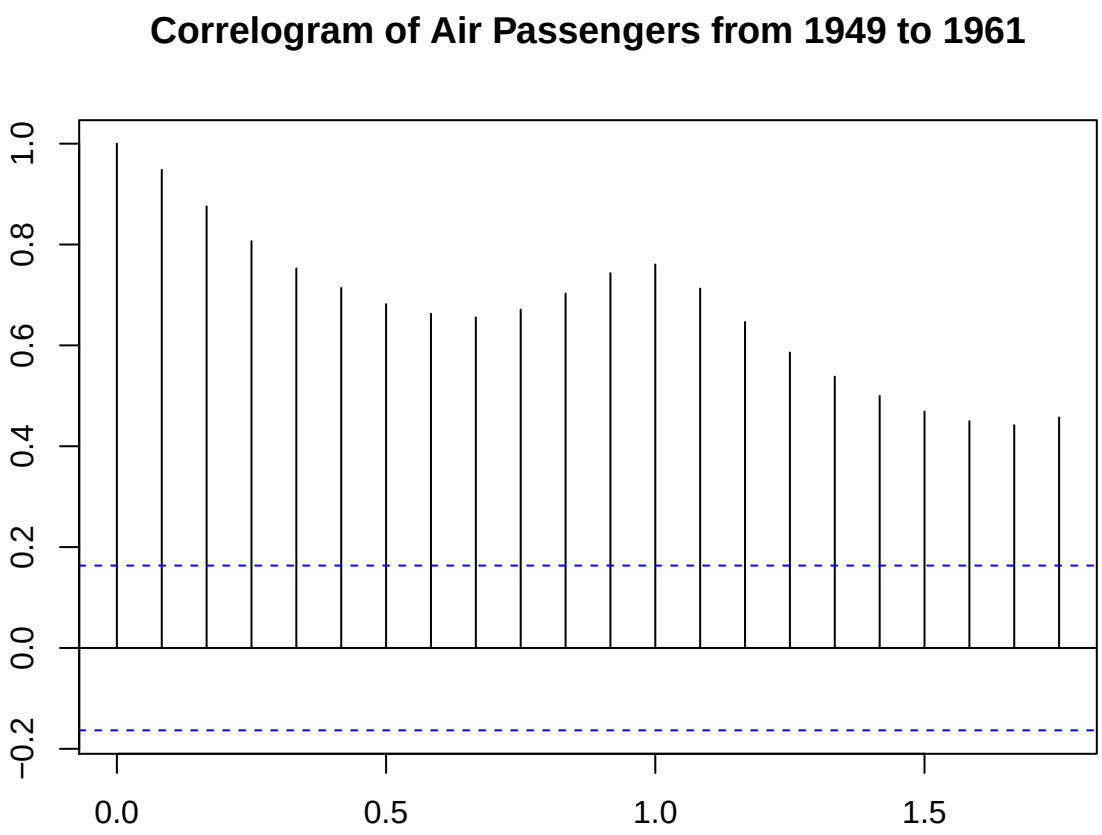


The passenger numbers increase over time with each year which may be indicative of an increasing linear trend, perhaps due to increasing demand for flight travel and commercialisation of airlines in that time period. In the boxplot there are more passengers travelling in months 6 to 9 with higher means and higher variances than the other months, indicating seasonality with a apparent cycle of 12 months. The rationale for this could be more people taking holidays and fly over the summer months in the US.

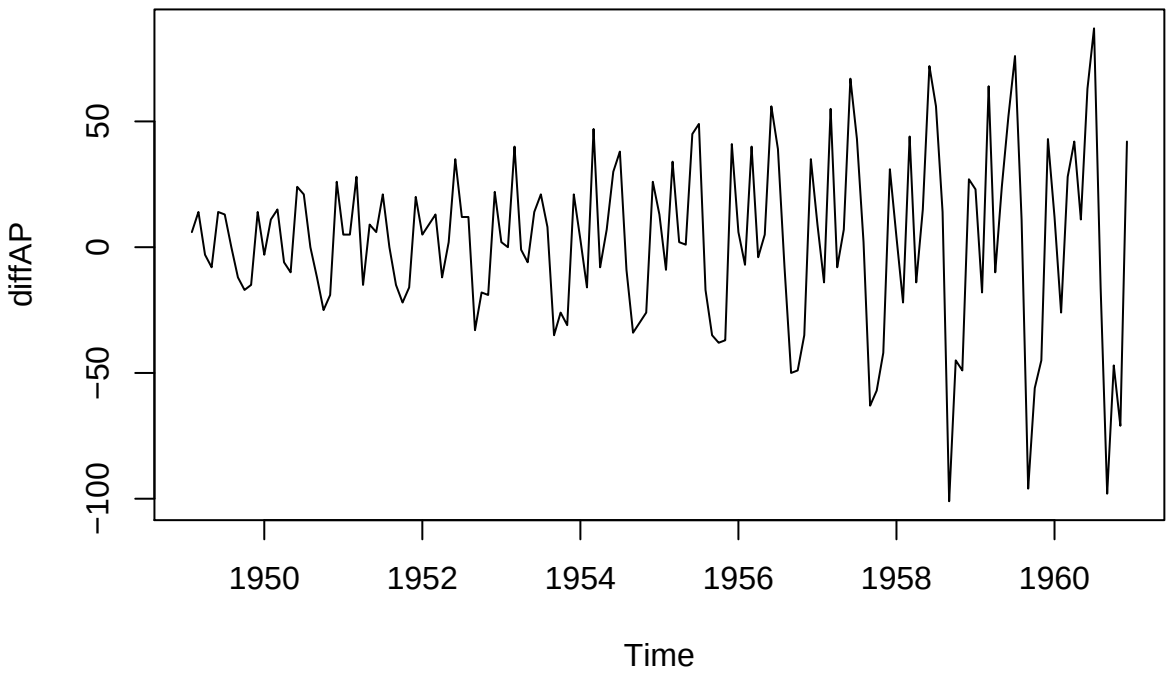
AirPassengers appears to be multiplicative time series as the passenger numbers increase, it appears so does the pattern of seasonality.

If we remove the trend from each series, we can reveal the "stationary" version of the data. Differencing is a very simple but very powerful technique for removing trend - it's just the difference between two neighboring points.

```
par(mar = c(3,3,3,3))
## Examine the auto-correlation function (ACF)
acf(AirPassengers, main = "Correlogram of Air Passengers from 1949 to 1961")
```

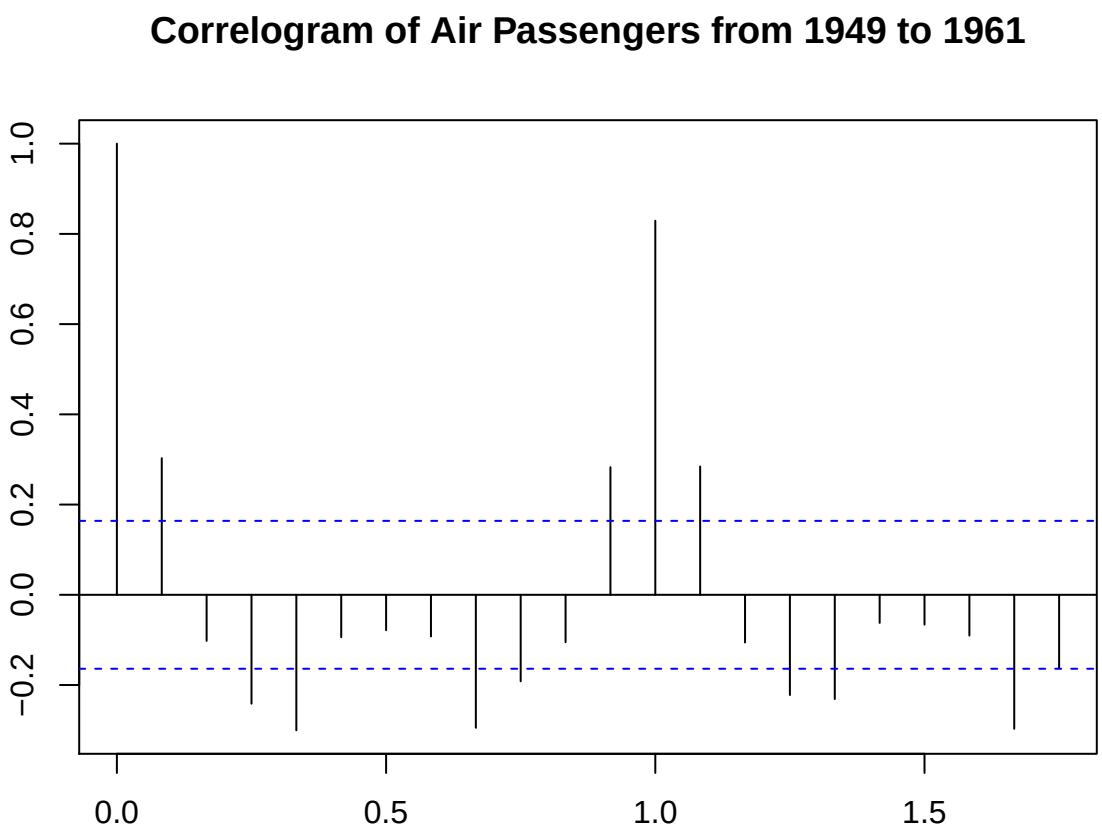



```
## Create a differenced data set
diffAP <- diff(AirPassengers)
plot(diffAP)
```



After differencing, the trend has been removed.

```
par(mar = c(3,3,3,3))
## Examine the auto-correlation function (ACF)
acf(diffAP, main = "Correlogram of Air Passengers from 1949 to 1961" ) # seasonalal
```



We can see that the acf of the residuals is centered around 0.

Dickey Fuller test for stationarity allows us to more definitively say, does this particular time series have any trend left in it. A significant DickeyFuller test means that a time series is stationary. We use the test after differencing(in our case) or decomposition to confirm that trend, seasonality, and other time artifacts have been removed.

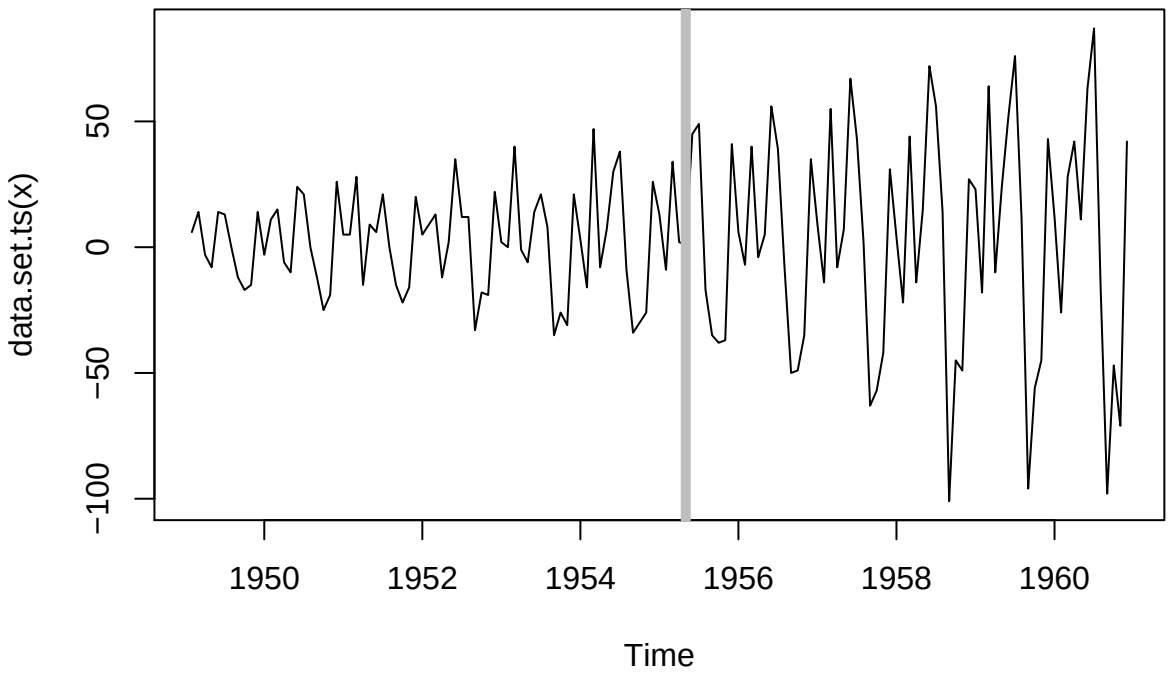
```
adf.test(diffAP) # shows significant, so it is stationary
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diffAP
## Dickey-Fuller = -7.0177, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

The output has a Dickey Fuller value in it(-7.01). It also shows a lag order that was the limit of the tests that were done and then a p-value(0.01), that is significant, and that means that the time series is stationary. If it's non-significant, then we have concerns that there may be trend in there or possibly other time series artifacts.

We can continue with further analysis of our time series.

```
diffAPcp <- cpt.var(diffAP)
plot(diffAPcp,cpt.col="grey",cpt.width=5)
```




```
diffAPcp # change in volatility occurred early in 1955
```

```
## Class 'cpt' : Changepoint Object
##         ~~   : S4 class containing 12 slots with names
##                cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 14.88853
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 76
```

In a change point analysis, an algorithm searches through the time series data to detect and document major transitions; provides one way of getting an insight into something important that may be happening in a time series. So change point analysis is useful to see if you can detect when some kind of event or intervention has occurred in a time series. This is a very useful way to do some research. We start measuring something at a time when we know not much is going on, then we keep measuring it through a period where there might be important changes happening. When we subject the data to analysis, we can notice whether or

not a meaningful change has occurred and that can be accomplished through a process called change point analysis.

cpt.var() finds the change point in the variability of the differenced time series (change in volatility occurred early in 1955).

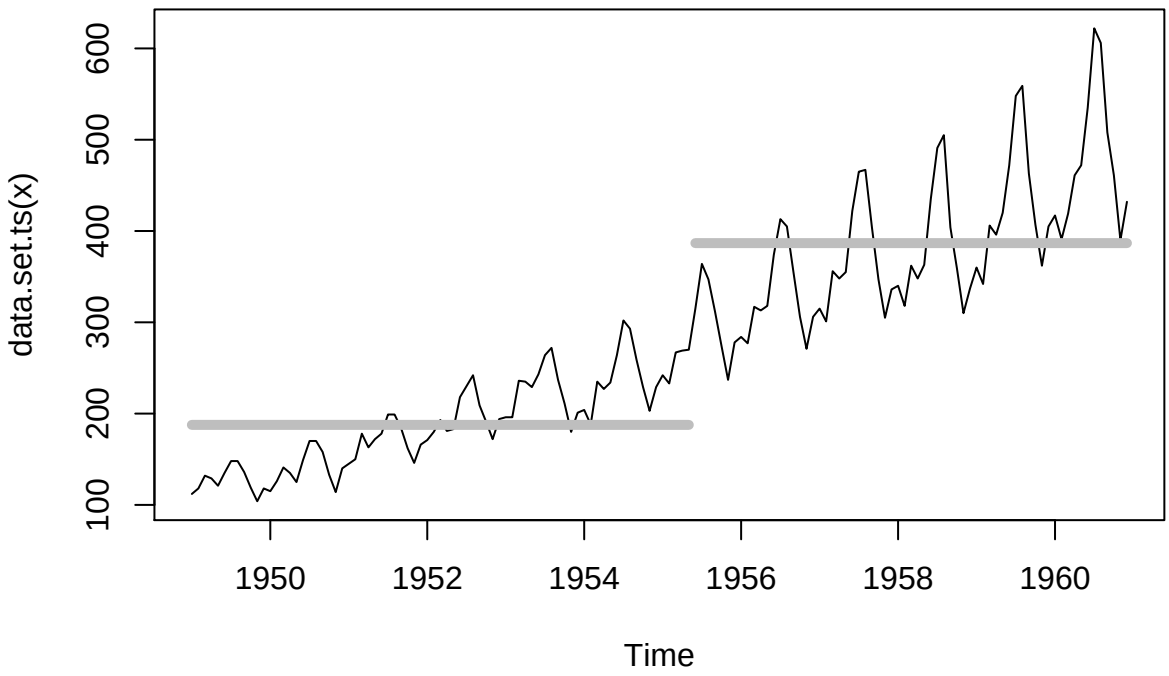
The major change point occurs at 76. This indicates that the amount of air passengers started to increase in 1955, and also entered a period of more intense volatility (grater variance). The "Type of Penalty" in the output refers to a mathematical formulation that determines how sensitive the algorithm is to detecting changes. MBIC = 14.88853 is like a statistical line, everything that crosses that like need to be examined.

6. Use cpt.mean() on the AirPassengers time series. Plot and interpret the results. Compare the change point of the mean that you uncovered in this case to the change point in the variance that you uncovered in Exercise 5. What do these change points suggest about the history of air travel?

```
diffAPcpM <- cpt.mean(AirPassengers)
diffAPcpM
```

```
## Class 'cpt' : Changepoint Object
##         ~~   : S4 class containing 12 slots with names
##                cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 14.90944
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 77
```

```
plot(diffAPcpM, cpt.col="grey",cpt.width=5)
```



The cpt.mean() function allows us to detect transition points where the mean of a time series changes substantively.

We have left the trend in the data. That is important because the change point analysis of means is usually concerned with places where a time series may have increased or decreased very suddenly.

The first few lines of the output of cpt.mean() simply document what the output object contains. When we get to "Method of Analysis," we learn that the algorithm has used "AMOC." This stands for "at most one change" and is confirmed a little later in the output with "Maximum no. of cpts : 1." In other words, we have asked the cpt.mean() procedure to look for one and only one shift in the mean of the time series. The "Type of Penalty" in the output refers to a mathematical formulation that determines how sensitive the algorithm is to detecting changes. cpt.mean() detected a change in the mean of the time series at point 77. The change in means is documented by the horizontal gray lines on the plot. The first line shows sort of the typical level, ranging from about 1949 to the beginning of 1955. There are some fluctuations in there, we do have a little bit of a growth pattern between 1950 and 1952, generally, they're all hanging out at that low level, which is an index level of about 200. Then in 1956 the index jumps up to a higher level, somewhere around 400. Each gray horizontal line represents the mean level of an index across the whole time period covered by the line. The change point is 77 (it was 76 with the var).

This result indicates that just as the air passangers started to increase rapidly in 1955, it also entered a period of more intense volatility, that is, substantially greater variance. The change points at 76 and 77 are part of the way through 1955.

```
diffAPcpM1 <- cpt.mean(AirPassengers, class = FALSE)
diffAPcpM1["conf.value"]
```

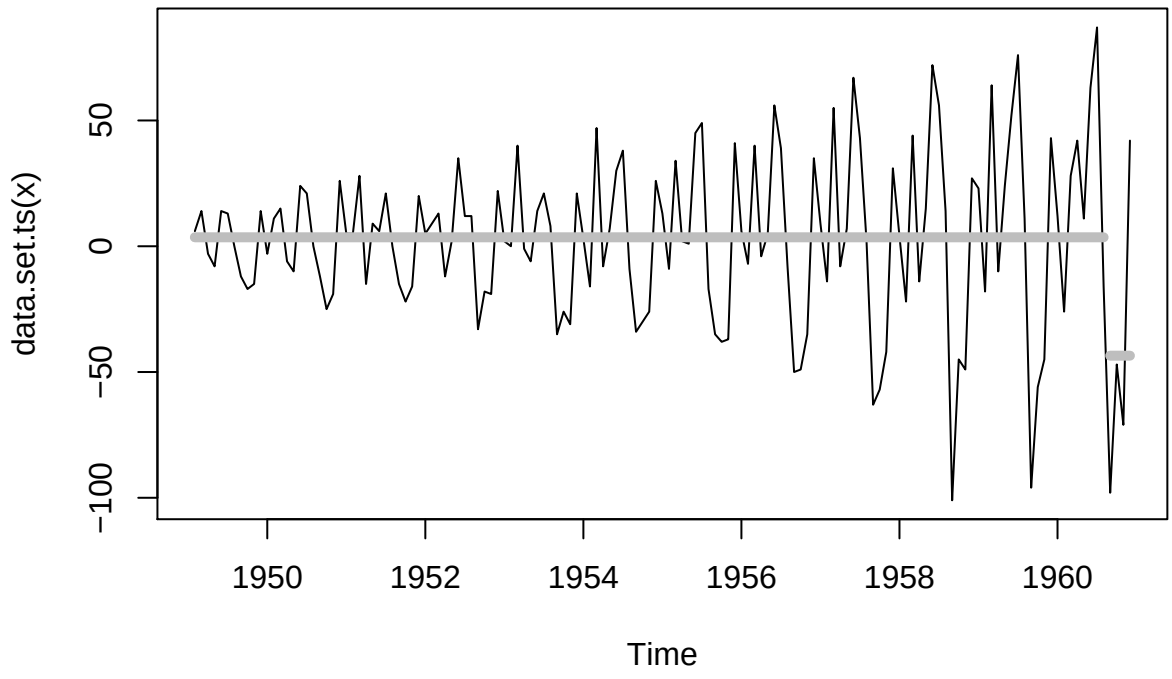
```
## conf.value
##          1
```

Confidence value is 1.0, the strongest possible value. This signifies that our analysis of the air passanger data has indeed detected a powerful change in the mean of the time series.

```
diffAPcpM2 <- cpt.mean(diffAP)
diffAPcpM2
```

```
## Class 'cpt' : Changepoint Object
##         ~~   : S4 class containing 12 slots with names
##                cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 14.88853
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 139
```

```
plot(diffAPcpM2, cpt.col="grey",cpt.width=5)
```



Removing the trends from the data eliminated the growth pattern, preventing us to find the shift in the means.

7. Find historical information about air travel on the Internet and/or in reference materials that sheds light on the results from Exercises 5 and 6. Write a mini- article (less than 250 words) that interprets your statistical findings from Exercises 5 and 6 in the context of the historical information you found.

The history of air travel began with the flight of the Wright brothers on December 17, 1903. Their plane was a powered and controlled aircraft whereas previous attempts to fly used gliders that had control and no power or free flight which had power but no control. After the First World War, as airplane designs became more reliable and planes larger in size, people and cargo began to be transported in aircraft. The beginning of World War II saw many airports being built in towns and cities with the availability of numerous qualified pilots. The history of air travel grew as light aircraft for the middle class market were developed by such growing manufacturers as Cessna, Piper and Beechcraft. In the history of air travel, civil air jets were developed by the 1950's with the first among them being the Boeing 707 passenger jet.

The 1945 invention of nuclear bombs briefly increased the strategic importance of military aircraft in the Cold War between East and West. Even a moderate fleet of long-range bombers could deliver a deadly blow to the enemy, so great efforts were made to develop countermeasures. At first, the supersonic interceptor aircraft were produced in considerable numbers. By 1955, most development efforts shifted to guided surface-to-air missiles. However, the approach diametrically changed when a new type of nuclear-carrying platform appeared that could not be stopped in any feasible way: intercontinental ballistic missiles. The possibility of these was demonstrated in 1957 with the launch of Sputnik 1 by the Soviet Union. This action started the Space Race between the nations.

Transcontinental schedules in the United States invariably included a stop for fuel en route; transatlantic flights between New York and Europe usually required refueling in Newfoundland, Iceland, or Ireland. These constraints began to evaporate in the 1950s with the Lockheed Super Constellation and the Douglas DC-7. The ultimate versions appeared in 1956–57 as the DC-7C, known as the "Seven Seas," which was capable of nonstop transatlantic flights in either direction, and the Lockheed 1649A Starliner, which could fly nonstop on polar routes from Los Angeles to Europe. The Starliner carried 75 passengers at speeds of 350 to 400 miles (560 to 640 km) per hour.

Although there were many benefits of flying in the 1950s and 1960s, the reality was far different than you might expect. In fact, once you know what flying during the so-called Golden Age was really like, you might prefer a jaunt on easyJet. Dangerous, smoky, boozy, boring, expensive, and racist is how the fights during that period of time was described(https://www.fastcompany.com/3022215/what-it-was-really-like-to-fly-during-the-golden-age-of-travel).

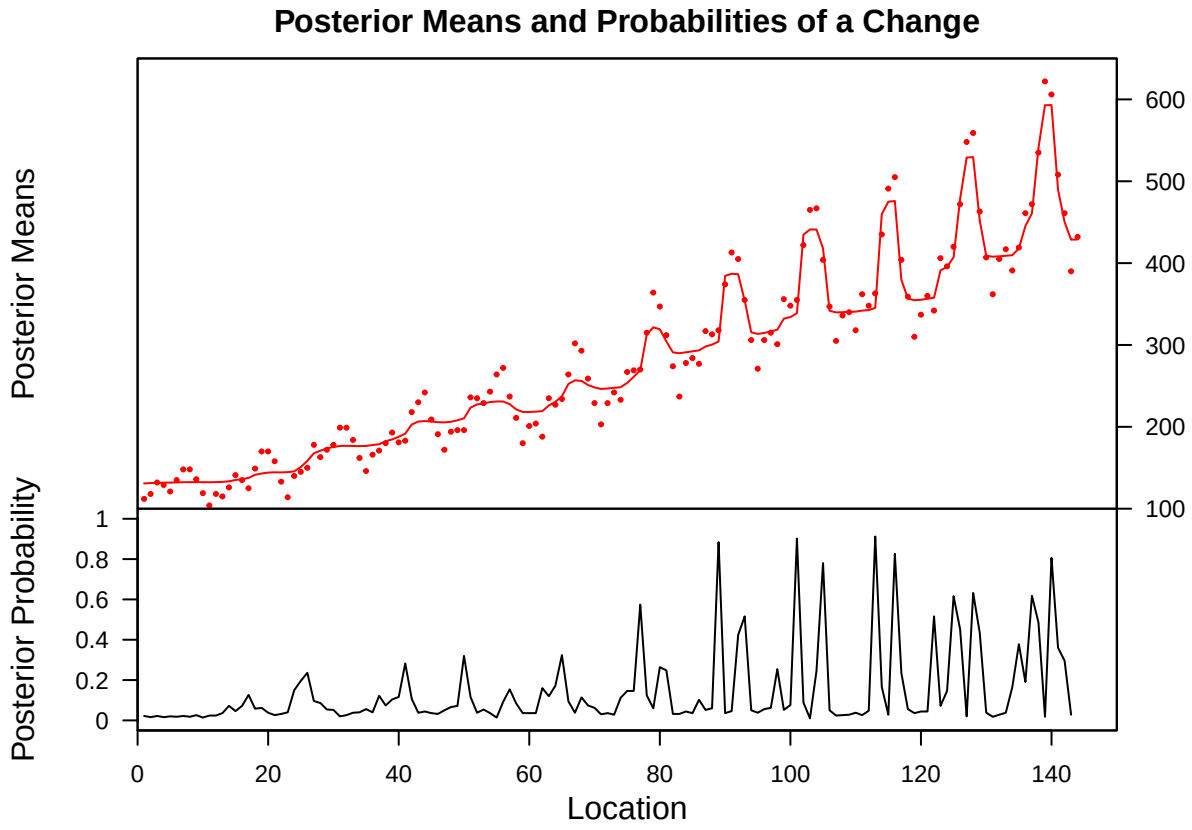
Airlines did encourage more people to fly in the 1950s and 1960s by introducing coach or tourist fares, but the savings were relative: less expensive than first class, but still pricey. In 1955, for example, so-called "bargain fares" from New York to Paris were the equivalent of just over $2,600 in 2014 dollars. Although the advent of jets did result in lower fares, the cost was still out of reach of most Americans. The most likely frequent flier was a white, male businessman traveling on his company's expense account, and in the 1960s, airlines – with young attractive stewardesses in short skirts – clearly catered to their most frequent flyers. The demographics of travelers did begin to shift during this period. More women, more young people, and retirees began to fly; still, airline travel remained financially out-of-reach for most.

We conducted an analysis of the air passangers over the years from 1949 to 1960. A strong positive trend was evident in the data. We conducted a change point analysis of means, using the "At Most One Change" (AMOC) search algorithm. A substantial change in the mean level of the index was detected in the early months of 1955. Likewise, after differencing to remove the trend, we conducted a change point analysis of variance with the AMOC algorithm and found a credible change in volatility at the same point in 1955. These changes in mean and variance appeared to coincide with the historical information about air travel.

8. Use bcp() on the AirPassengers time series. Plot and interpret the results. Make sure to contrast these results with those from Exercise 6.

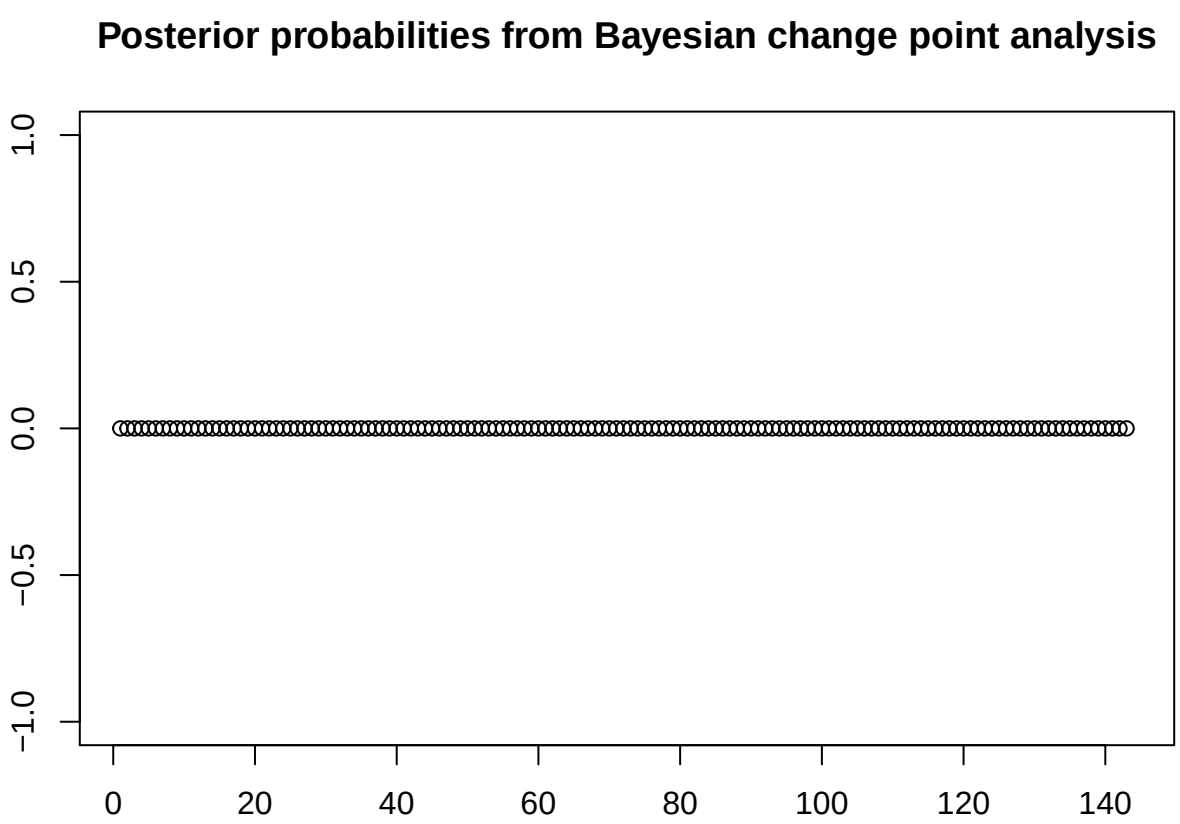
This Bayesian version of change point analysis uses the Markov chain Monte Carlo technique to develop a list of posterior priorities for mean changes at each point in the time series.

```
bcpdiffAP <- bcp(as.vector(AirPassengers))
plot(bcpdiffAP, outer.margins = list(left = unit(4,"lines"), bottom = unit(3, "lines"), right = unit(3,
```



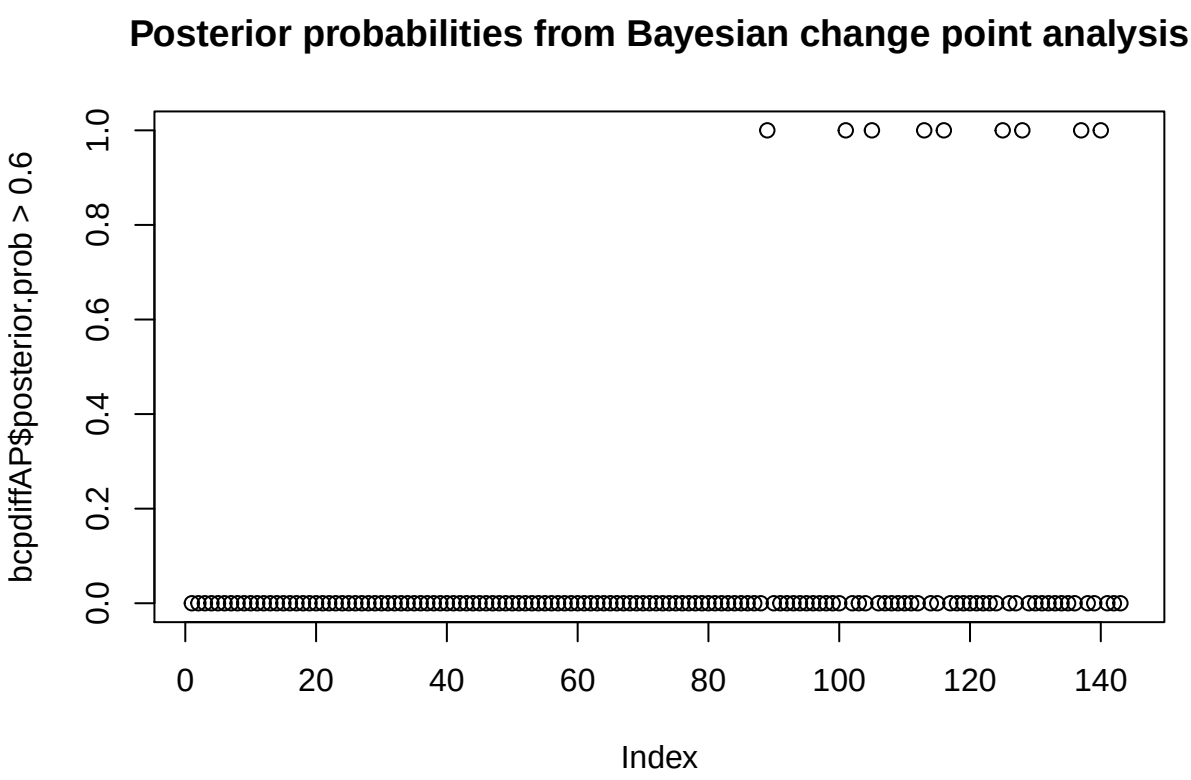
The upper pane shows the original time series and the lower pane shows the probabilities of a mean change at each point in time. Rather than trying to identify one particular point where a change may have occurred, what the Bayesian method does is it shows you the probabilities that a change occurred. There are isolated spikes that show probabilities near 1 at one point across the timeline. Somewhere near data point 77 we see there is some density of probability.

```
set.seed(1234)
# dev.off()
par(mar = c(2,2,4,2))
## Reploting the probabilities
plot(bcpdiffAP$posterior.prob >.95, main = "Posterior probabilities from Bayesian change point analysis
```



Every probability value less than or equal to 0.95 gets recoded as FALSE (which in R is equivalent to 0) whereas everything above 0.95 gets coded as TRUE (which in R is equivalent to 1). In our case we have all but one FALSE. Earlier we saw established change points, here the probabilities are below .95, except only one point that is around 117 with a very strong probability of being change point(it was 77 with the mean change point).

```
## Reploting the probabilities
plot(bcpdiffAP$posterior.prob >.60, main = "Posterior probabilities from Bayesian change point analysis
```



When we set the threshold lower, we get more change points. The plot shows that there are one such points just above 80, another two points just above 100 and 200, and then another point just above 140 with .6 probabilities of being change points - signifying a substantial and sustained change in the mean of the time series. These change points correspond to periods just before 1949 and just after 1952, contrasting the results from earlier. With the monotonic rise in this data set, the choice of a single change point for the mean is going to be somewhat subjective.