Time Series Data and Random Samples

Time series data are different from a simple random sample.

Consider these data:

112 118 132 129 121 135 148 148 136 119 104 118
115 126 141 135 125 149 170 170 158 133 114 140
145 150 178 163 172 178 199 199 184 162 146 166
171 180 193 181 183 218 230 242 209 191 172 194
196 196 236 235 229 243 264 272 237 211 180 201
204 188 235 227 234 264 302 293 259 229 203 229
242 233 267 269 270 315 364 347 312 274 237 278
284 277 317 313 318 374 413 405 355 306 271 306
315 301 356 348 355 422 465 467 404 347 305 336
340 318 362 348 363 435 491 505 404 359 310 337
360 342 406 396 420 472 548 559 463 407 362 405
417 391 419 461 472 535 622 606 508 461 390 432

Look at some graphs and summary statistics.
summary(aa)

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>104.0</td>
<td>180.0</td>
<td>265.5</td>
<td>280.3</td>
<td>360.5</td>
<td>622.0</td>
</tr>
</tbody>
</table>
Time Series Data

Now consider the same data with some metadata.

<table>
<thead>
<tr>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1949</td>
<td>112</td>
<td>118</td>
<td>132</td>
<td>129</td>
<td>121</td>
<td>135</td>
<td>148</td>
<td>148</td>
<td>136</td>
<td>119</td>
<td>104</td>
</tr>
<tr>
<td>1950</td>
<td>115</td>
<td>126</td>
<td>141</td>
<td>135</td>
<td>125</td>
<td>149</td>
<td>170</td>
<td>170</td>
<td>158</td>
<td>133</td>
<td>114</td>
</tr>
<tr>
<td>1951</td>
<td>145</td>
<td>150</td>
<td>178</td>
<td>163</td>
<td>172</td>
<td>178</td>
<td>199</td>
<td>199</td>
<td>184</td>
<td>162</td>
<td>146</td>
</tr>
<tr>
<td>1952</td>
<td>171</td>
<td>180</td>
<td>193</td>
<td>181</td>
<td>183</td>
<td>218</td>
<td>230</td>
<td>242</td>
<td>209</td>
<td>191</td>
<td>172</td>
</tr>
<tr>
<td>1953</td>
<td>196</td>
<td>196</td>
<td>236</td>
<td>235</td>
<td>229</td>
<td>243</td>
<td>264</td>
<td>272</td>
<td>237</td>
<td>211</td>
<td>180</td>
</tr>
<tr>
<td>1954</td>
<td>204</td>
<td>188</td>
<td>235</td>
<td>227</td>
<td>234</td>
<td>264</td>
<td>302</td>
<td>293</td>
<td>259</td>
<td>229</td>
<td>203</td>
</tr>
<tr>
<td>1955</td>
<td>242</td>
<td>233</td>
<td>267</td>
<td>269</td>
<td>270</td>
<td>315</td>
<td>364</td>
<td>437</td>
<td>312</td>
<td>274</td>
<td>237</td>
</tr>
<tr>
<td>1956</td>
<td>284</td>
<td>277</td>
<td>317</td>
<td>313</td>
<td>318</td>
<td>374</td>
<td>413</td>
<td>405</td>
<td>355</td>
<td>306</td>
<td>271</td>
</tr>
<tr>
<td>1957</td>
<td>315</td>
<td>301</td>
<td>356</td>
<td>348</td>
<td>355</td>
<td>422</td>
<td>465</td>
<td>467</td>
<td>404</td>
<td>347</td>
<td>305</td>
</tr>
<tr>
<td>1958</td>
<td>340</td>
<td>318</td>
<td>362</td>
<td>348</td>
<td>363</td>
<td>435</td>
<td>491</td>
<td>505</td>
<td>404</td>
<td>359</td>
<td>310</td>
</tr>
<tr>
<td>1959</td>
<td>360</td>
<td>342</td>
<td>406</td>
<td>396</td>
<td>420</td>
<td>472</td>
<td>548</td>
<td>559</td>
<td>463</td>
<td>407</td>
<td>362</td>
</tr>
<tr>
<td>1960</td>
<td>417</td>
<td>391</td>
<td>419</td>
<td>461</td>
<td>472</td>
<td>535</td>
<td>622</td>
<td>606</td>
<td>508</td>
<td>461</td>
<td>390</td>
</tr>
</tbody>
</table>

Let’s plot it again, this time with lines and proper labeling of the axes. This is all done automatically by R, once it is told this is a time series and what the relevant times are.
Time Series Data

> class(AirPassengers)
[1] "ts"
> start(AirPassengers); end(AirPassengers); frequency(AirPassengers)
[1] 1949 1
[1] 1960 12
[1] 12
> plot(AirPassengers, ylab="Passengers (1000’s)")
Time Series Data

Now we see some things that might be interesting and might help us understand the data

** and may help us understand the data-generating process that gave rise to the data.

It might also help us to forecast future values.

We first notice a **trend**.

It is more or less linear, but linear regression would not make much sense.

Because of the trend, however, simple statistics such as mean and variance or other things given by summary would not make much sense.
Cycles

As we mentioned, although there is a linear trend, linear regression would not make much sense.

We notice cycles in the data.

“Cycle” is a generic term that we will use for this kind of pattern in the data.

Looking more closely, and think about what the source of the data, we see that the cycles have fixed periods.

We call such cycles seasonal effects, where the “season” may be any fixed unit of time, quarters, months, weeks, days, hours, etc.

There are many other “cycles”, such as business cycles, cycles of the Southern Oscillation, etc. These may not have fixed periods.
Meanings of Cycles

When we see a seasonal variation, we should think about the nature of the data-generating process.

In the airline passenger data, the seasonal variation makes sense.

This was expected when the R ts dataset was built. Remember \texttt{frequency(AirPassengers)}.

In the case of the airline passenger data, the seasons are months. We could have set it up differently.

In R, \texttt{aggregate} can remove the seasonal effect by summing over the seasons, and \texttt{cycle} forms groups.

\begin{verbatim}
layout(1:2)
plot(aggregate(AirPassengers), ylab="Passengers (1000’s")
boxplot(AirPassengers ~ cycle(AirPassengers))
\end{verbatim}
<table>
<thead>
<tr>
<th>Time</th>
<th>Passengers (1000's)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950</td>
<td>100</td>
</tr>
<tr>
<td>1952</td>
<td>200</td>
</tr>
<tr>
<td>1954</td>
<td>300</td>
</tr>
<tr>
<td>1956</td>
<td>400</td>
</tr>
<tr>
<td>1958</td>
<td>500</td>
</tr>
<tr>
<td>1960</td>
<td>600</td>
</tr>
</tbody>
</table>
**Time Series Data**

So if simple statistics such as mean and variance do not make much sense, what should we do?

In a simple random sample of realizations of a random variable $X$, we have $x_1, x_2, \ldots, x_n$. There is no real difference in $x_1$ and $x_2$, except for sampling variation.

In a time series, the random variable itself has an index: $X_t$.

We could talk about the mean of $X_t$ at any fixed $t$, $E(X_t)$.

It could be the case that we have a sample of observations at each time point, $x_{t1}, x_{t2}, \ldots, x_{tn}$, and for a fixed $t$ there would be no real difference in $x_{t1}$ and $x_{t2}$, except for sampling variation.

In most time series data, we only have one observation on the random variable $X_t$ at any fixed $t$. 
Time Series Data

In some cases, we may be interested in a mean of a subsequence or in how the mean changes over time: at time $t$,

$$E(X_t) = \mu_t$$

*** notation!!

- mean of a random walk is the starting point

- mean of a random walk with drift is a function of time, it is the starting point plus the total drift at that point in time.

- mean of signal plus noise is a function of time, it is the value of the signal at that point in time.
Time Series Data

Instead of an overall mean or variance, we are generally interested in how the data changes over time.

- **autocovariance function (ACVF):**
  \[ \gamma_x(s, t) = \mathbb{E} \left( (x_s - \mu_s)(x_t - \mu_t) \right) \]

- **autocorrelation function (ACF):**
  \[ \rho_x(s, t) = \frac{\gamma_x(s, t)}{\sqrt{\gamma_x(s, s)\gamma_x(t, t)}} \]

The autocovariance and the autocorrelation of white noise is 0 if \( s \neq t \).

The autocovariance of a moving average of a white noise process is 0 only outside the window of the moving average.
Bivariate Time Series

Often a time series consists of bivariate data at each time point:

\[ x_1, x_2, \ldots, x_n \]
\[ y_1, y_2, \ldots, y_n \]

In this case we are generally interested in how the two series change together.

- **cross-covariance function (CCVF):**
  \[
  \gamma_{xy}(s, t) = \mathbb{E}((x_s - \mu_x)(y_t - \mu_y))
  \]

- **cross-correlation function (CCF):**
  \[
  \rho_{xy}(s, t) = \frac{\gamma_{xy}(s, t)}{\sqrt{\gamma_x(s, s)\gamma_y(t, t)}}
  \]
Stationarity

Stationarity is an important property of some time series.

There are various levels of stationarity.

- **strictly stationary** time series:
  The joint probability distribution of any set of points

  \[ \{x_{t_1}, \ldots, x_{t_k}\} \]

  is the same as that of the set

  \[ \{x_{t_1}+h, \ldots, x_{t_k}+h\} \]

  Strict stationarity is too restrictive to be of much use.
Stationarity

- **weakly stationary** time series:
  - The variance at each point is finite.
  - The mean function $\mu_t$ is constant (that is, it does not depend on $t$).
  - The autocovariance function $\gamma(s, t)$ is constant for any fixed difference $|t - s|$.

Because strict stationarity is too restrictive to be of much use, when we just say “stationary” we will mean weakly stationary.

A strictly stationary process is clearly weakly stationary.
Stationarity

Because autocovariance function $\gamma(s, t)$ of a stationary time series depends only on the difference $|t - s|$, we often write $s = t + h$, for some given value $h$, and so write the autocovariance function as $\gamma(h)$ and the autocorrelation function $\rho(h)$.

We should note that another notation is in common usage. Instead of “$\gamma_x(h)$” and “$\rho_x(h)$”, some people write “$\gamma_h(x)$” and “$\rho_h(x)$”.

A white noise process is strictly stationary (show this).

A moving average of a white noise process is (weakly) stationary.
Joint Stationarity

A bivariate process \( \{x_t, y_t\} \) is said to be **jointly stationary** if each process is stationary and the cross-covariance function \( \gamma_{xy}(s, t) \) is constant for fixed values of \( |t - s| \); that is, for \( h = s - t \)

\[
\gamma_{xy}(s, t) = \gamma_{xy}(h) = \mathbb{E} \left( (x_{t+h} - \mu_x)(y_t - \mu_y) \right).
\]

We should note that some people define \( \gamma_{xy}(h) \) as

\[
\mathbb{E} \left( (x_t - \mu_x)(y_{t+h} - \mu_y) \right).
\]

We will use the notation as we have defined it above. This is the interpretation used in the R functions. In any event, note that

\[
\gamma_{xy}(h) = \gamma_{yx}(-h).
\]

We define the **cross-correlation function (CCF)** of a jointly stationary process to be

\[
\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}.
\]
Statistical Modeling

Statistical models identify observational data with realizations of a probability distribution.

The objective generally is to develop a model in which some component of the observational data has a common (and simple) distribution over all individual observations.

\[ Y_i = f(i, \{y_j : j < i\}, x_i, \theta) + \epsilon_i \]

All of the components, \( Y_i, i, \{y_j : j < i\}, x_i \) are observable except \( f, \theta, \) and \( \epsilon_i \).

We assume \( f \) are \theta constant, and \( \epsilon_i \) has an iid probability distribution.