정상 확률 과정과 비정상 확률 과정

정상 확률 과정

협의의 정상 확률 과정(strictly stationary process, strong stationary process)은 확률 과정의 모든 모멘트(moment)가 시간 차이(time lag)에만 의존하고 절대 시간에 의존하지 않는 것이다.

이를 수학적으로 표현하면 임의의 t, s, k_i에 대해

$$\text{E}[Y_t Y_{t+k_1} Y_{t+k_2} \cdots Y_{t+k_i} \cdots] = \text{E}[Y_s Y_{s+k_1} Y_{s+k_2} \cdots Y_{s+k_i} \cdots]$$

가 성립한다.

즉. 기댓값의 경우

$$\text{E}[Y_t] = \text{E}[Y_s] = \mu$$

가 성립하고

자기공분산의 경우

$$\text{E}[Y_t Y_{t+k}] = \text{E}[Y_s Y_{s+k}] = f(k)$$

가 성립한다.

위 두가지 조건만 성립하는 경우에는 **광의의 정상 확률 과정(wide-sense stationary process, weak stationary process)**라고 한다.

정상 확률과정에서는 자기공분산의 값이 시간 변수의 차이 즉 시차(lag) k에만 의존한다. 따라서 자기공분산은 시차에 대한 함수이다. 이를 **자기공분산함수(auto covariance function)**라고 한다.

$$\gamma_{t,t+k} = \gamma_{0,k} \triangleq \gamma_k$$

정상 확률 과정의 자기상관계수도 마찬가지로 시차 k에만 의존한다. 이를 **자기상관계수 함수(auto-correlation function)** 줄여서 ACF라고 한다.

$$\rho_{t,t+k} = \rho_{0,k} \triangleq \rho_k = \frac{\gamma_k}{\gamma_0}$$

정상 확률 과정은 다음과 같은 성질을 만족한다.

$$\begin{aligned} \gamma_0 &= \text{Var}[Y_t] \\ \gamma_k &= \gamma_{-k} \\ |\gamma_k| &\leq \gamma_0 \\ \rho_0 &= 1 \\ \rho_k &= \rho_{-k} \\ |\rho_k| &\leq 1 \end{aligned}$$

에르고딕 성질

정상 확률 과정에서는 각각의 시간에 해당하는 확률 변수의 무조건부 분포가 모두 같다. 따라서 시계열 데이터를 이루는 각 숫자가 하나의 분포에서 나온 표본 데이터 집합이라고 생각할 수 있다. 이 성질을 이용하면 기댓값이나 자기공분산 등에서 필요한 앙상블 평균을 계산할 때 여러개의 시계열 데이터 표본이 필요하지 않고 단 하나의 시계열 데이터 표본만 있어도 된다.

기댓값의 경우,

$$\lim_{N\to\infty} \frac{1}{N} \sum^{N} Y_t = \mathrm{E}[Y_t]$$

자기공분산의 경우

$$\lim_{N\to\infty} \frac{1}{N} \sum^{N} Y_t Y_{t+k} = \mathrm{E}[Y_t Y_{t+k}]$$

가 성립한다.

이를 에르고딕 성질(ergodicity)이라고 한다.

비정상 확률 과정

정상 확률 과정이 아닌 확률 과정이 **비정상 확률 과정(non-stationary process)**이다.

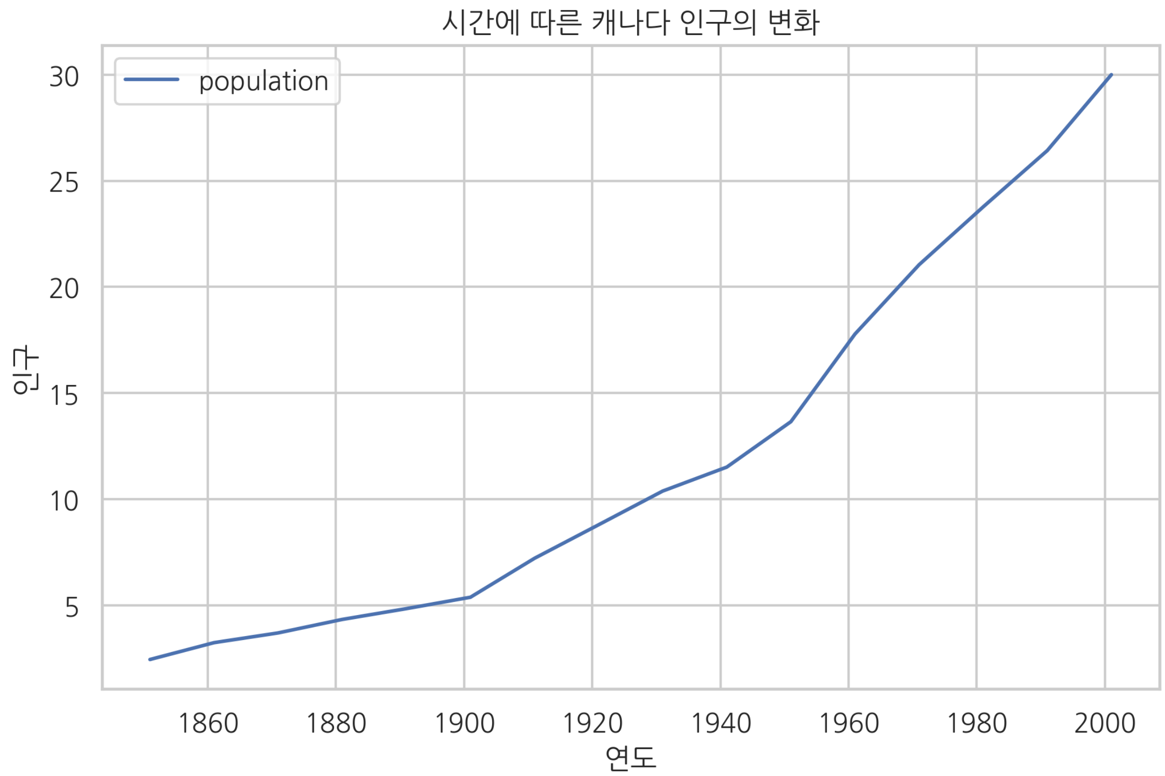
비정상 확률 과정이 되는 경우는

- 추세를 가지는 경우. 일차 모멘트 즉, 기댓값 $\mathrm{E}[y_t]$이 시간에 따라 변화함
- 추세가 없지만 분산 $\mathrm{Var}[y_t]$ 이 시간에 따라 변하는 경우

등이 있다.

In [1]:

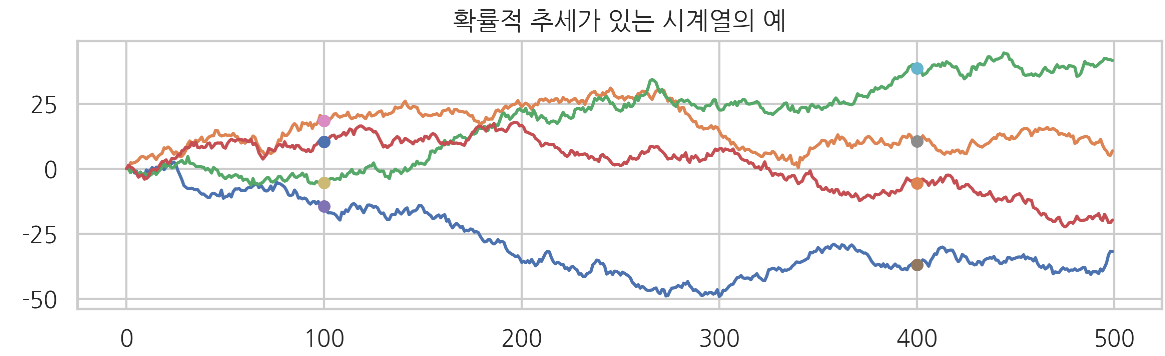
```
df = sm.datasets.get_rdataset("CanPop", package="carData").data
df.plot(x="year", y="population")
plt.xlabel("연도")
plt.ylabel("인구")
plt.title("시간에 따른 캐나다 인구의 변화")
plt.show()
```

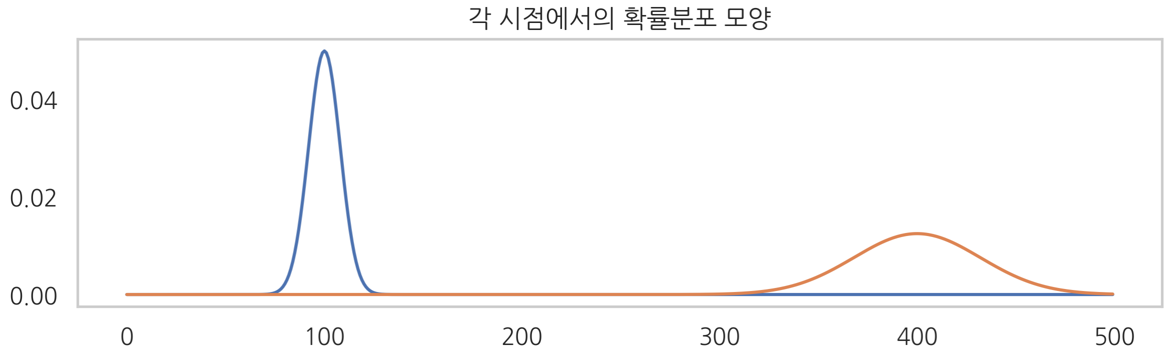


다음 시계열 자료들은 동일한 확률 과정의 샘플들이다. 하나 하나의 샘플(시계열 자료)만 보면 마치 추세가 있는 것처럼 보인다. 그러나 이는 확률 과정의 분산 $\text{Var}[y_t]$이 시간 t에 따라 커지기 때문이다. 그래프에 표시하였듯이 $t = 400$에서의 분산 $\text{Var}[y_{400}]$은 $t = 100$에서의 분산 $\text{Var}[y_{100}]$보다 크다. 이런 경우 **확률적 추세(stochastic trend)**를 가진다고 말하기도 한다.

In [2]:

```
N = 500
t1 = 100
t2 = 400
t = np.arange(N)

np.random.seed(12)
y1 = np.insert(np.cumsum(sp.stats.norm.rvs(size=N-1)), 0, 0)
np.random.seed(18)
y2 = np.insert(np.cumsum(sp.stats.norm.rvs(size=N-1)), 0, 0)
np.random.seed(22)
y3 = np.insert(np.cumsum(sp.stats.norm.rvs(size=N-1)), 0, 0)
np.random.seed(24)
y4 = np.insert(np.cumsum(sp.stats.norm.rvs(size=N-1)), 0, 0)

plt.subplot(211)
plt.title("확률적 추세가 있는 시계열의 예")
plt.plot(t, y1)
plt.plot(t, y2)
plt.plot(t, y3)
plt.plot(t, y4)
plt.plot(t1, y1[t1], 'o', markersize=5)
plt.plot(t2, y1[t2], 'o', markersize=5)
plt.plot(t1, y2[t1], 'o', markersize=5)
plt.plot(t2, y2[t2], 'o', markersize=5)
plt.plot(t1, y3[t1], 'o', markersize=5)
plt.plot(t2, y3[t2], 'o', markersize=5)
plt.plot(t1, y4[t1], 'o', markersize=5)
plt.plot(t2, y4[t2], 'o', markersize=5)

plt.subplot(212)
plt.grid(False)
plt.title("각 시점에서의 확률분포 모양")
plt.plot(t, sp.stats.norm(t1, 0.08*t1).pdf(t))
plt.plot(t, sp.stats.norm(t2, 0.08*t2).pdf(t))

plt.tight_layout()
plt.show()
```


확률적 추세가 있는 시계열의 예
25
0
-25
-50
0
100
200
300
400
500


각 시점에서의 확률분포 모양
0.04
0.02
0.00
0
100
200
300
400
500