Explanation of
`seasonal_decompose` (Statsmodels)

1. Overview

Given a univariate time series $\{y_t\}_{t=1}^{N}$, the classical decomposition implemented by `statsmodels.tsa.seasonal.se`
separates y_t into three components:

- **Trend** component: T_t
- **Seasonal** component: S_t
- **Residual** component: R_t

Depending on the chosen *model* argument, the decomposition can be either additive:

$$y_t \;=\; T_t \;+\; S_t \;+\; R_t\,, \tag{1}$$

or multiplicative:

$$y_t \;=\; T_t \;\times\; S_t \;\times\; R_t\,. \tag{2}$$

We denote by P the *period* (number of observations per cycle). For example, if working with hourly data and expecting a 24-hour (daily) cycle, then $P = 24$.

2. Trend Estimation

2.1. Moving-Average Filter

To estimate the trend component $\{T_t\}$, the algorithm applies a *centered moving average* of length P. Let $m = \lfloor (P-1)/2 \rfloor$. If P is odd, the trend at time t is computed as

$$\widehat{T}_t \;=\; \frac{1}{P} \sum_{i=-m}^{m} y_{\,t+i}\,.$$

If P is even, a two-stage filter is used (a length-P moving average followed by a length-2 average) so that the resulting filter weights are symmetric and centered. In effect:

$$\widehat{T}_t \;=\; \frac{1}{2}\Big[\underbrace{\tfrac{1}{P} \sum_{i=-(P/2-1)}^{P/2} y_{\,t+i}}_{\text{length-}P\text{ MA}} \;+\; \underbrace{\tfrac{1}{2}\Big(y_{t-P/2} + y_{t+P/2}\Big)}_{\text{length-2 MA}}\Big].$$

In either case, endpoints $t \le m$ or $t > N - m$ cannot form a full centered window, so $\widehat{T}_t$ is set to NaN for those indices.

2.2. Detrending

Once we have the raw trend estimate $\widehat{T}_t$, we remove it from the original series to obtain a *detrended* series $\{D_t\}$:

- **Additive model:**
$$D_t \;=\; y_t \;-\; \widehat{T}_t\,.$$

- **Multiplicative model:**
$$D_t \;=\; \frac{y_t}{\widehat{T}_t}\,.$$

By construction, $\{D_t\}$ still contains the seasonal oscillations of period P plus residual noise.

3. Seasonal Estimation

3.1. Seasonal Phase-Averaging (Expanded Discussion)

The key assumption in classical decomposition is that the seasonal component $\{S_t\}$ repeats exactly every P time steps. To estimate $\{S_t\}$, the algorithm performs *phase-averaging*, also called *seasonal averaging*, which proceeds as follows:

3.1.1. Defining Phases. First, assign each time index t to a "phase" within the cycle:

$$\text{phase}(t) \;=\; (t \bmod P), \quad \text{where } \text{phase}(t) \in \{\,0, 1, \dots, P-1\}.$$

In practice, if the input time series $\{y_t\}$ is equally spaced, index t can be the integer position in the array (e.g., $t = 0, 1, \dots, N-1$). For example:

- If $P = 24$ on hourly data, "phase" 0 might represent midnight–1am, phase 1 represents 1am–2am, and so on.
- If $P = 96$ on 15-minute bars for a daily cycle, phase 0 is 00:00–00:15, phase 1 is 00:15–00:30, ..., phase 95 is 23:45–24:00.

3.1.2. Grouping by Phase. For each phase index $i \in \{0, 1, \dots, P-1\}$, collect all detrended values D_t such that $\text{phase}(t) = i$. Denote

$$\mathcal{I}_i \;=\; \{\,t : \, t \bmod P = i\}, \qquad N_i = |\mathcal{I}_i| \quad \text{(number of points in phase } i\text{)}.$$

Then form the raw seasonal estimate for that phase by taking the arithmetic mean (additive model) or geometric mean (multiplicative model) of the detrended values at those positions.

- **Additive phase-mean:**
$$\widetilde{S}_i \;=\; \frac{1}{N_i} \sum_{t \in \mathcal{I}_i} D_t \;=\; \frac{1}{N_i} \sum_{\substack{t=0 \\ t \bmod P = i}}^{N-1} (\, y_t - \widehat{T}_t\,)\,. \tag{3}$$

- **Multiplicative phase-mean:**
$$\widetilde{S}_i \;=\; \Bigl(\prod_{t \in \mathcal{I}_i} D_t\Bigr)^{1/N_i} \;=\; \exp\Bigl(\frac{1}{N_i} \sum_{t \in \mathcal{I}_i} \ln\bigl(\tfrac{y_t}{\widehat{T}_t}\bigr)\Bigr). \tag{4}$$

3.1.3. Why Phase-Averaging?

- Over multiple cycles, each phase i should exhibit the same underlying seasonal effect plus noise. By averaging across all cycles, random fluctuations (noise) tend to cancel out, leaving the *persistent* seasonal pattern at that phase.
- If the time series spans M full cycles (so $N \approx M \cdot P$), then N_i is approximately M for each i.
  - In practice, if N is not an exact multiple of P, some phases will have $\lfloor N/P \rfloor$ points, others $\lceil N/P \rceil$. The algorithm simply uses whatever count arises naturally.
  - Endpoints (first $\lfloor P/2 \rfloor$ and last $\lfloor P/2 \rfloor$ points) may lack corresponding trend estimates, so such t are omitted from D_t (and thus from phase means). This can slightly adjust N_i at the margins, but if $N \gg P$, edge effects are negligible.
- Phase-averaging implicitly assumes *stationary seasonality*—that is, the seasonal shape does not systematically change from one cycle to the next. If a seasonal pattern drifts over time (e.g., gradually shifting peak times), classical phase-averaging may blur those effects into a less sharp average.

3.1.4. Normalization of Phase Means. Once the raw phase-means $\{\widetilde{S}_0, \widetilde{S}_1, \dots, \widetilde{S}_{P-1}\}$ have been computed via (3) or (4), they are *normalized* so that the seasonal component has zero sum (additive) or unit product (multiplicative) over one full cycle.

- **Additive normalization:**

$$\overline{S} \;=\; \frac{1}{P}\sum_{i=0}^{P-1} \widetilde{S}_i, \qquad \widehat{S}_i \;=\; \widetilde{S}_i - \overline{S}, \quad i = 0, \dots, P-1.$$

  This ensures

$$\sum_{i=0}^{P-1} \widehat{S}_i = 0.$$

- **Multiplicative normalization:**

$$G \;=\; \exp\!\Big(\tfrac{1}{P}\sum_{i=0}^{P-1} \ln(\widetilde{S}_i)\Big), \quad \widehat{S}_i = \frac{\widetilde{S}_i}{G}, \quad i = 0, \dots, P-1.$$

  This enforces

$$\prod_{i=0}^{P-1} \widehat{S}_i = 1, \quad \text{equivalently} \quad \frac{1}{P}\sum_{i=0}^{P-1} \ln(\widehat{S}_i) = 0.$$

3.1.5. Constructing the Full Seasonal Series. After normalization, the algorithm extends the vector $\{\widehat{S}_0, \dots, \widehat{S}_{P-1}\}$ to a length-N sequence $\{\widehat{S}_t\}$ by:

$$\widehat{S}_t = \widehat{S}_{t \bmod P}, \quad t = 0, 1, \dots, N-1.$$

Thus, each index t in the original series is assigned the seasonal value corresponding to its phase. The result is a repeated "template" of length P tiled across the entire time axis.

3.1.6. Handling Missing or Irregular Timestamps.

- The classical implementation requires a *regularly spaced* index so that each cycle has exactly P positions (ignoring endpoints). If the input data have missing timestamps (e.g., weekends missing in daily data), one typically *resamples* to a fixed frequency and either forward-fills or interpolates gaps before decomposition.
- If some phases have fewer observations due to missing data, the corresponding $\widetilde{S}_i$ will be averaged over fewer points. When N is large, this distortion is often minor, but if many consecutive timestamps are missing (e.g., a holiday), it can bias certain phase means. A more robust approach would downweight those phases or impute missing D_t before averaging.

3.1.7. Example Illustration. Suppose you have 10 days of hourly data ($P = 24$, $N = 240$). Then:

- Phase 0 corresponds to all timestamps at midnight. You'll collect D_t for $t \in \{0, 24, 48, \ldots, 216\}$ (10 points) and average them to get $\widetilde{S}_0$.
- Phase 1 corresponds to 1am–2am, so $t \in \{1, 25, 49, \ldots, 217\}$. Again average those 10 points for $\widetilde{S}_1$, and so on.
- After computing all $\widetilde{S}_i$, adjust by subtracting the mean $\overline{S}$ so that $\sum_{i=0}^{23} \widehat{S}_i = 0$. Then set $\widehat{S}_t = \widehat{S}_{t \bmod 24}$ for each $t = 0, \ldots, 239$.

3.1.8. Interpretation of Seasonal Profiles.

- The resulting sequence $\{\widehat{S}_t\}$ reveals how y_t systematically deviates from its trend at each position within the cycle.
- Plotting one cycle (e.g., $\{\widehat{S}_0, \ldots, \widehat{S}_{P-1}\}$) against "time-within-cycle" produces a template that can expose:
  - Peak and trough times (e.g., intraday hours of high or low volume/price).
  - Asymmetry or skew in the pattern (e.g., slow rise, rapid fall).
  - Periods of relative stability (phases where $\widehat{S}_i \approx 0$).
- If the series has multiple nested seasonalities (e.g., daily and weekly), classical phase-averaging only captures one specified P. To detect multiple cycles, one can either:
  1. Perform a two-stage approach (first remove daily seasonality, then average residuals over a weekly period), or
  2. Use more advanced methods (e.g., STL or TBATS) that allow simultaneous extraction of multiple seasonal components.

3.2. Normalization

To ensure that the seasonal component has mean zero (for additive) or geometric mean one (for multiplicative) over each cycle, we apply a normalization step:

- **Additive model:** Force
$$\sum_{i=0}^{P-1} \widehat{S}_i \;=\; 0.$$
Concretely, let
$$\overline{S} \;=\; \frac{1}{P}\sum_{i=0}^{P-1} \widetilde{S}_i, \qquad \widehat{S}_i \;=\; \widetilde{S}_i \;-\; \overline{S}, \quad i = 0, \ldots, P-1.$$
Then extend $\widehat{S}_i$ to length N by $\widehat{S}_t = \widehat{S}_{t \bmod P}$.

- **Multiplicative model:** Force
$$\prod_{i=0}^{P-1} \widehat{S}_i \;=\; 1.$$
Equivalently, ensure $\displaystyle\frac{1}{P}\sum_{i=0}^{P-1} \log(\widehat{S}_i) = 0$. Concretely, let
$$G = \exp\Bigl(\tfrac{1}{P}\sum_{i=0}^{P-1} \ln(\widetilde{S}_i)\Bigr), \quad \widehat{S}_i = \frac{\widetilde{S}_i}{G}, \quad i = 0, \ldots, P-1,$$
and extend by $\widehat{S}_t = \widehat{S}_{t \bmod P}$.

4. Residual Calculation

With $\widehat{T}_t$ and $\widehat{S}_t$ in hand, the residual component $\{R_t\}$ is simply:

- **Additive:**
$$\widehat{R}_t \;=\; y_t \;-\; \widehat{T}_t \;-\; \widehat{S}_t. \tag{5}$$

- **Multiplicative:**
$$\widehat{R}_t \;=\; \frac{y_t}{\widehat{T}_t \;\times\; \widehat{S}_t}. \tag{6}$$

Again, for t in the first or last $\lfloor P/2 \rfloor$ indices, $\widehat{T}_t$ (and thus $\widehat{S}_t$) may be undefined (NaN), so $\widehat{R}_t$ is also NaN there.

5. Summary of Algorithm

Putting it all together, the algorithmic steps for

```
decomp = seasonal_decompose(y, model='additive', period=P)
```

are:

1. **Trend Smoothing.**

$$\widehat{T}_t \;=\; \text{Centered-Moving-Average}(y_t, \text{window} = P).$$

   Endpoints where a full centered window cannot be computed are set to NaN.

2. **Detrending.**

$$D_t = y_t - \widehat{T}_t \quad \text{(additive)}, \quad \text{or} \quad D_t = \frac{y_t}{\widehat{T}_t} \quad \text{(multiplicative)}.$$

3. **Seasonal Phase-Averaging.**

   - For each phase $i = 0, 1, \ldots, P-1$, compute

$$\widetilde{S}_i = \begin{cases} \dfrac{1}{N_i} \displaystyle\sum_{t\,:\,t \bmod P = i} \bigl(y_t - \widehat{T}_t\bigr), & \text{(additive)} \\ \exp\Bigl(\frac{1}{N_i} \displaystyle\sum_{t\,:\,t \bmod P = i} \ln\bigl(\frac{y_t}{\widehat{T}_t}\bigr)\Bigr), & \text{(multiplicative)} \end{cases}$$

   - Normalize $\{\widetilde{S}_i\}$ so that

$$\sum_{i=0}^{P-1} \widehat{S}_i = 0 \quad \text{(additive)}, \quad \text{or} \quad \prod_{i=0}^{P-1} \widehat{S}_i = 1 \quad \text{(multiplicative)}.$$

   - Extend to all $t = 1, \ldots, N$ by $\widehat{S}_t = \widehat{S}_{t \bmod P}$.

4. **Residuals.**

$$\widehat{R}_t = y_t - \widehat{T}_t - \widehat{S}_t \quad \text{(additive)}, \quad \text{or} \quad \widehat{R}_t = \frac{y_t}{\widehat{T}_t\,\widehat{S}_t} \quad \text{(multiplicative)}.$$

5. **Output.** The result returned by `seasonal_decompose` contains four series (each length N with NaNs at endpoints):

$$\Bigl\{\widehat{y}_t = y_t,\ \widehat{T}_t,\ \widehat{S}_t,\ \widehat{R}_t\Bigr\}, \quad t = 1, \ldots, N.$$

6. Important Notes

- **Selection of Period P.** The user must specify P, the number of observations per seasonal cycle. A correct P is crucial. For instance:
  - *Hourly data, daily cycle:* $P = 24$.
  - *Daily data, weekly cycle:* $P = 7$.
  - *5-minute data, daily cycle:* $P = 24 \times 12 = 288$.

  If P is mis-specified, the estimated seasonal component will not correspond to a real repeating pattern.

- **Additive vs. Multiplicative.**
  - *Additive* (Equation 1): use when seasonal fluctuations are roughly constant in absolute terms.
  - *Multiplicative* (Equation 2): use when seasonal amplitude scales with the overall level (e.g. seasonal swings are proportionally larger when y_t is larger).

- **Endpoint NaNs.** For $t \le \lfloor P/2 \rfloor$ or $t > N - \lfloor P/2 \rfloor$, the trend cannot be fully computed (lack of a centered window). As a result:

  $$\widehat{T}_t = \text{NaN}, \quad \widehat{S}_t = \text{NaN}, \quad \widehat{R}_t = \text{NaN}.$$

  Only the interior $(P/2)$–offset region has valid estimates.

- **Interpretation of Residuals.** If the chosen trend and seasonal components capture the systematic structure well, the residuals $\{\widehat{R}_t\}$ should resemble "white noise" (no obvious autocorrelation or pattern). Large spikes in $\widehat{R}_t$ indicate times when y_t deviated sharply from both its long-run path and its typical repeating pattern.

References:

- Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice.* OTexts.
- `statsmodels` documentation: `https://www.statsmodels.org/{stable}/generated/statsmodels.tsa.seasonal.seasonal_decompose.html`