9.1 Measurements of time series

Contents

(All variables are real and one-dimensional unless otherwise specified.)

Suppose we have obtained a time series that is guaranteed to become **stationary** after some time. As the properties of a stationary time series no longer changes with time, we can estimate its **population** mean and variance with their **sample** counterparts reasonably.

1. Revisiting mean and variance

As you have surely known, the mean μ and variance σ^2 of a continuous random variable X are fundamentally defined with its distribution $f(x)$.

$$\mu \equiv \int x f(x) \mathrm{d}x$$

$$\sigma^2 \equiv \int (x-\mu)^2 f(x) \mathrm{d}x$$

These are, precisely speaking, the **population mean** and **population variance** of X because they are constants and do not depend on any specific realizations. Note that

$$\begin{aligned}
\sigma^2 &= \int (x-\mu)^2 f(x) \mathrm{d}x \\
&= \underbrace{\int x^2 f(x) \mathrm{d}x}_{\langle X^2 \rangle} - 2\mu \underbrace{\int x f(x) \mathrm{d}x}_{\mu} + \mu^2 \underbrace{\int f(x) \mathrm{d}x}_{1} \\
&= \langle X^2 \rangle - \mu^2 \, .
\end{aligned}$$

The mean tells us the **expected value** of X, i.e. literally, the average value of X you expect to see after almost infinitely many trials, but the expected value itself may never be realized at all. (The expected value of a fair dice is 3.5, but you will never get this value with the dice.) The variance measures the **dispersion** of X around its mean, aka its **central tendency**.

However, we rarely know the distribution and thus the values of population mean and population variance, so we need to **estimate them with samples** instead. Given a set of realizations $\{x_1, x_2, \dots, x_n\}$, the **sample mean** $\hat{\mu}$ of X is simply

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \,,$$

whereas its **sample variance** $\hat{\sigma}^2$ is defined as

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \equiv \frac{n}{n-1} \left(\overline{x^2} - \hat{\mu}^2 \right)$$

for $\overline{x^2} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i^2$. Although there are n terms in the summation of $\hat{\sigma}^2$, the estimate is argued to be more accurate if $\frac{1}{n-1}$ is used as the leading factor instead of $\frac{1}{n}$. This is known as **Bessel's correction**.

1.1 Notations

The mentioned concepts may be denoted by different symbols in different contexts **inconsistently**.

- **Letters.** Population mean and population variance are commonly denoted by Greek letters μ and σ^2. Their sample counterparts may be denoted by Roman letters m and s^2 instead.
- **Bar.** A bar ¯ is commonly put on a variable's sample mean, so $\bar{x}$ may denote the mean of $\{x_1, x_2, \dots, x_n\}$, i.e. the sample mean of X.
- **Hat.** Statistics commonly puts a hat ˆ on an estimated value, so $\hat{\mu}$ may denote the estimated value of μ.
- **Angle brackets.** Physics commonly encloses a population mean (also called an **ensemble average** in physics) with angle brackets $\langle\rangle$, so $\langle X \rangle$ may denote the population mean of X.
- **Operators**. An explicit operator E may be used to emphasize that a population mean represents an expected value, so $E(X)$ may denote the population mean of X.

1.2 Central limit theorem

The sample mean $\hat{\mu}$ is used to estimate the population mean μ, so we would like to know how much $\hat{\mu}$ differs from μ. But before answering this, we need to understand that $\hat{\mu}$ is also a random variable since we obtain different samples means from different trials of the same experiment. The distribution of $\hat{\mu}$ may be called the **sampling distribution of mean.**

While each realization of $\hat{\mu}$ is calculated by averaging $\{x_1, x_2, \dots, x_n\}$, the **central limit theorem** states that as the **sample size** n approaches infinity, the sample distribution becomes a

normal distribution with mean μ and variance $\frac{\sigma^2}{n}$, where μ and σ^2 are the population mean and population variance of X,

$$\hat{\mu} \overset{n\to\infty}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

1.3 Standard error of mean

We can therefore quantify how much the sample mean $\hat{\mu}$ of a specific trial differs from the population mean μ by the **standard deviation** $\sqrt{\frac{\sigma^2}{n}}$ of the sampling distribution (given that n is large enough). But since we do not know the population variance σ^2, we need to replace this standard deviation with another estimate

$$\delta = \sqrt{\frac{\hat{\sigma}^2}{n}},$$

where $\hat{\sigma}^2$ is the sample variance of the trial. The value δ is called the **standard error of mean**.

Qualitatively, a small δ suggests a high accuracy in $\hat{\mu}$, so the accuracy increases as the sample size n grows. Quantitatively, **frequentist statistics** tells us that the probability that $\hat{\mu}$ **covers** μ follows the standard normal distribution, i.e.

$$P\left(\hat{\mu} - z\delta \le \mu \le \hat{\mu} + z\delta\right) = C(z) \equiv \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-x^2/2} \mathrm{d}x$$

for some $z > 0$. The range $[\hat{\mu} - z\delta, \hat{\mu} + z\delta]$ is called the $[100\, C(z)]\,\%$ **confidence interval** of μ.

Interpretation. Frequentists argue that $[\hat{\mu} - z\delta, \hat{\mu} + z\delta]$ is a **variable range** because $\hat{\mu}$ and δ are random variables yet to be measured. Once they are determined to be $\hat{\mu}_{\{x\}}$ and $\delta_{\{x\}}$ by a particular set of realizations $\{x\}$, the range collapses into a **fixed range** $\left[\hat{\mu}_{\{x\}} - z\delta_{\{x\}}, \hat{\mu}_{\{x\}} + z\delta_{\{x\}}\right]$ and implies

$$P\left(\hat{\mu}_{\{x\}} - z\delta_{\{x\}} \le \mu \le \hat{\mu}_{\{x\}} + z\delta_{\{x\}}\right) \in \{0, 1\}$$

because the population mean μ is a constant that certainly falls into or outside a fixed range. For example, given $z = 1.96 \Rightarrow C(z) \approx 0.95$ and thus 95% confidence intervals, frequentists never claim the proposition "$\mu \in \left[\hat{\mu}_{\{x\}} - 1.96\,\delta_{\{x\}}, \hat{\mu}_{\{x\}} + 1.96\,\delta_{\{x\}}\right]$" is 95% true—it is either 100% true or 100% false—but they only claim that after repeating the experiment for many times, 95% of all the obtained ranges $\left\{\left[\hat{\mu}_{\{x\}} - 1.96\,\delta_{\{x\}}, \hat{\mu}_{\{x\}} + 1.96\,\delta_{\{x\}}\right]\right\}$ satisfy the proposition "$\mu \in [\hat{\mu} - 1.96\,\delta, \hat{\mu} + 1.96\,\delta]$".

Unfortunately, it is very common for scientists to use these frequentist concepts without following the frequentist interpretation.

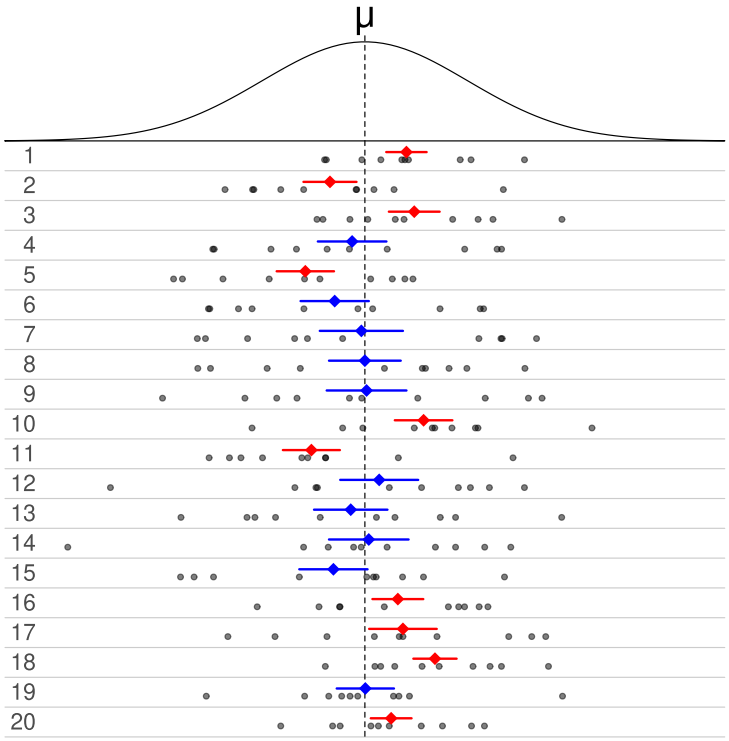

Fig. 1. This picture summarizes the previous paragraphs. Retrieved from [Wikimedia Commons (https://upload.wikimedia.org/wikipedia/commons/5/5c/Normal_distribution_50%25_CI_illustration.svg)](https://upload.wikimedia.org/wikipedia/commons/5/5c/Normal_distribution_50%25_CI_illustration.svg).

- The bell curve shows the distribution of the sample mean $\hat{\mu}$, i.e. the sampling distribution of mean, of a random variable X. The sample means $\{\hat{\mu}\}$ are normally distributed around X's population mean μ once they are calculated with a large sample size.
- 20 trials are performed to measure X. Each set of realizations $\{x\}$ (the black dots) yield a sample mean $\hat{\mu}_{\{x\}}$ (the red or blue dots) and a standard error $\delta_{\{x\}}$ and thus form a fixed range $\left[\hat{\mu}_{\{x\}} - z\delta_{\{x\}}, \hat{\mu}_{\{x\}} + z\delta_{\{x\}}\right]$ (the horizontal bars). Here, $C(z) = 0.5 \Rightarrow z = 0.67$ is used, so the ranges are 50% confidence intervals.
- Some of the fixed ranges can cover μ (the blue dots and bars), but some cannot (the red ones). Since they are 50% confidence intervals, 50% of the trials are expected to produce a range that can cover μ.

Now, X is no longer a random variable but a time series $X = \{X_t \mid 1 \leq t \leq T\}$, which is random but surely stationary after some time. When we are asked to calculate its sample mean $\hat{\mu}$

and sample variance $\hat{\sigma}^2$, we may be tempted to treat the values $\{X_1, X_2, \ldots, X_T\}$ like the realizations $\{x_1, x_2, \ldots, x_n\}$ of a random variable and thus claim

$$\hat{\mu} = \frac{1}{T}\sum_{i=1}^{T} X_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{T-1}\sum_{i=1}^{T} (X_i - \hat{\mu})^2 \,.$$

However, this approach is problematic because unlike $\{x_1, x_2, \ldots, x_n\}$, $\{X_1, X_2, \ldots, X_T\}$ lack **statistical independence** and yield biased results.

2. Autocorrelation and independence

We can extract statistically independent data points from $\{X_1, X_2, \ldots, X_T\}$ for the calculation by considering its **autocorrelation** A, which tells us how strongly X correlates with its previous self. As we learnt previously, A is defined as

$$A_\lambda = \langle (X_t - \mu)\,(X_{t+\lambda} - \mu) \rangle$$

for some discrete lag $\lambda \geq 0$, where μ is the population mean of X and the ultimate quantity that we are looking for. Doesn't this create a circular dependency?

2.1 Practical definitions

There are several ways to realize the formula in order to to resolve the problem. Here are two possibilities:

$$\begin{aligned} A_\lambda^{(1)} &= \left\langle \left(X_t - \frac{1}{T}\sum_{t'=1}^{T} X_{t'} \right) \left(X_{t+\lambda} - \frac{1}{T}\sum_{t'=1}^{T} X_{t'} \right) \right\rangle \\ &= \frac{1}{T}\sum_{t=1}^{T-\lambda} X_t X_{t+\lambda} - \left(\frac{1}{T}\sum_{t=1}^{T} X_t \right)^2 \end{aligned}$$

and

$$\begin{aligned} A_\lambda^{(2)} &= \left\langle \left(X_t - \frac{1}{T-\lambda}\sum_{t'=1}^{T-\lambda} X_{t'} \right) \left(X_{t+\lambda} - \frac{1}{T-\lambda}\sum_{t'=1}^{T-\lambda} X_{t'+\lambda} \right) \right\rangle \\ &= \frac{1}{T-\lambda}\sum_{t=1}^{T-\lambda} X_t X_{t+\lambda} - \left(\frac{1}{T-\lambda}\sum_{t=1}^{T-\lambda} X_t \right) \left(\frac{1}{T-\lambda}\sum_{t=1}^{T-\lambda} X_{t+\lambda} \right) \,. \end{aligned}$$

Other definitions exist and differ in the prior estimate of μ and the normalization constant, i.e. $\dfrac{1}{T}$ versus $\dfrac{1}{T-\lambda}$.

I prefer the first one because it is simpler in terms of programming, so we will use it in the following sections. Let us rewrite its expression by defining $\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$.

$$A_\lambda = A_\lambda^{(1)} = \frac{1}{T}\sum_{t=1}^{T-\tau} X_t X_{t+\lambda} - \bar{X}^2$$

2.2 Fourier analysis

While the autocorrelation of a continuous-time series can be easily derived with a **Fourier transform**, some special steps are, however, required before applying it on a discrete-time series X. To do so, let $Y = \{Y_t \mid 1 \le t \le 2T\}$ be a time series with

$$Y_t = \begin{cases} X_t - \bar{X} & (1 \le t \le T) \\ 0 & (T < t \le 2T) \end{cases} .$$

It is **padded with zeros** because X is not a periodic signal required by a Fourier transform. Then, we can compute another $2T$-unit long time series $Z = \{Z_t \mid 1 \le t \le 2T\}$ with

$$Z = \frac{1}{T}\mathcal{F}^{-1}\left[\left|\mathcal{F}(Y)\right|^2\right],$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ respectively denote Fourier transform and its inverse. [Remember that $\mathcal{F}(Y)$ is a complex object, so $\left|\mathcal{F}(Y)\right|^2$ refers to the product between $\mathcal{F}(Y)$ and its complex conjugate.] The autocorrelation of X is finally given by

$$A_\lambda^{(3)} = Z_{\lambda+1} \quad \text{for} \quad \lambda \in [0, T-1] \ ;$$

surprisingly, $A_\lambda^{(1)} \equiv A_\lambda^{(3)}$ without any numerical differences, and this implies $A_\lambda^{(1)}$ is more accurate than $A_\lambda^{(2)}$.

Technically speaking, $\mathcal{F}$ and $\mathcal{F}^{-1}$ are performed by the algorithm of **fast Fourier transform** (FFT). It can speed up the computation of autocorrelation from $O\left(T^2\right)$, which $A_\lambda^{(1)}$ gives, to $O(T \ln T)$. This is the greatest (and perhaps the only) advantage of using $A_\lambda^{(3)}$. If you do not understand how to program an FFT, you should always stick to $A_\lambda^{(1)}$ because—as the saying goes—"people's time is more expensive than computers' time".

2.3 Correlation time

Once we obtain A, we can determine at least how far two data points $\left(X_{t_1}, X_{t_2}\right)$ should be separated so that they can be regarded as statistically independent. The answer is related to its characteristic timescale, namely **correlation time** τ. If X is a **Markov process**, we can estimate τ by fitting

$$A_\lambda \sim e^{-\lambda/\tau} \, .$$

A more sophisticated analysis may fit

$$A_\lambda \sim e^{-(\lambda/\alpha)^\beta}$$

to get $\tau = \dfrac{\alpha}{\beta} \, \Gamma\!\left(\dfrac{1}{\beta} \right)$ instead, but this practice is rare and does not seem to produce better results.

Integrated correlation time. If the autocorrelation is perfectly exponential, i.e. $A_\lambda = A_0 e^{-\lambda/\tau}$, we may estimate $\tau \equiv \displaystyle\int_0^\infty e^{-\lambda/\tau} \, \mathrm{d}\lambda$ by **numerically integrating** $\dfrac{A_\lambda}{A_0}$. This alternative estimate may be called the "integrated correlation time". Some people argue that this method is more accurate than fitting, but I doubt this personally.

2.4 Interpretation

Autocorrelation may be viewed as a measure of **self-similarity**. While X_t and $X_{t+\tau}$ are on average $e^{-\tau/\tau} \approx 37\%$ similar, the similarity between X_t and $X_{t+2\tau}$ drops to $e^{-2\tau/\tau} \approx 14\%$, which happens to be a sufficiently good threshold of statistical independence. Consequently, $X_{1+2\tau}$ is deemed to have "forgotten" X_1, so we may even assume that X starts to "forget" its history and thus become stationary at $t = 2\tau$.

But of course, the arguments here may not be universally true because they rely on numerous assumptions on the underlying form of X. You should modify them wisely if, for example, X apparently does not look stationary after $t = 2\tau$ or the autocorrelation A does not resemble any exponential decay at all.

3. Sample statistics

We have now extracted $n \equiv \lfloor T/\tau \rfloor$ independent data points $\{X_{2\tau}, X_{4\tau}, X_{6\tau}, \ldots\}$ from X. Its sample mean $\hat{\mu}$ is then refined to become

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_{2\tau i} \, ,$$

whereas its sample variance $\hat{\sigma}^2$ accordingly reads

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_{2\tau i} - \hat{\mu} \right)^2 .$$

3.1 Standard error of mean

The standard error of the sample mean $\hat{\mu}$ is given by

$$\delta = \sqrt{\frac{1}{n}\,\hat{\sigma}^2}\,.$$

If you somehow do not care about or bother to compute the sample variance $\hat{\sigma}^2$, you may alternatively compute δ with

$$\delta = \sqrt{\frac{1+2\tau}{T}\,\hat{\sigma}_0^2}\,,$$

where $\hat{\sigma}_0^2 = \dfrac{1}{T-1}\sum_{t=1}^{T}\left(X_t - \bar{X}\right)^2$ is the sample variance suggested by the original unfiltered X. While they should be identical in theory, they may differ considerably in reality.

3.2 Standard error of variance?

We would also like to know how much the sample variance $\hat{\sigma}^2$ differs from the population variance, but we can hardly write down a formula for its standard error because it is affected by the error of $\hat{\mu}$ via some tedious calculus. Instead, we can make our lives easier by measuring the error algorithmically with the **bootstrap method**.

1. Let $B = \{X_{2\tau}, X_{4\tau}, X_{6\tau}, \ldots\}$ be the n independent data points, then denote their sample variance by s_1.
2. Draw $|B| = n$ terms from B **with replacement** and record the sample variance s_2 of these new terms.
3. Repeat step 2 as many times possible to gather many sample variances $\{s_1, s_2, s_3, \ldots\}$.
4. After sufficient trials, the **sample standard deviation** of $\{s_1, s_2, s_3, \ldots\}$ will converge to a value, and it is the standard error of $\hat{\sigma}^2$.

In fact, the method does not require statistically independent data, so you may replace $B = \{X_{2\tau}, X_{4\tau}, X_{6\tau}, \ldots\}$ with $B = \{X_1, X_2, \ldots, X_T\}$ in step 1, but the expense is that $|B| = n$ will increase to $|B| = T$ in step 2 and thus slow down the algorithm.

If you want to better understand the principles of all the methods mentioned here, you may read Sections 3.3 and 3.4 of *Monte Carlo Methods in Statistical Physics* (Barkema and Newman, 1999). Although it is written in physicists' language, the sections explain some useful techniques for handling a stationary time series.