Non Normality and Non Stationarity

by Kumar Shantanu

> "I can live with doubt and uncertainty and not knowing. I think it's much more interesting to live not knowing than to have answers that might be wrong." — Richard P. Feynman, The Pleasure of Finding Things Out: The Best Short Works of Richard P. Feynman

Non-Normality

Technical Definition

A random variable say $X \in \mathbb{R}$ is said to be normally distributed if its probability density function is given by (assuming $\mathbb{R}$ is the support of X):

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ if } X \in \mathbb{R} \text{ or } 0 \text{ otherwise}$$

where μ is the mean and σ is the standard deviation of the distribution.

A theoretical normal distribution is characterized by two parameters: the mean μ and the standard deviation σ. The mean is the center of the distribution, and the standard deviation is a measure of how spread out numbers are. The standard deviation is the square root of the variance, which is the average of the squared deviations from the mean. The variance is a measure of how far a set of (random) numbers are spread out from their average value.

Why normal distribution is a big deal?

I think there are two reasons why normal distribution is a big deal (probably more, but I can't think of them right now):

1. **Asymptotic property of many estimators:** Many estimators of population parameters are asymptotically normal. This means that as the sample size increases, the distribution of the estimator gets closer and closer to a normal distribution. Most famous of them is the Central Limit Theorem (CLT) which states that the distribution of any linear combination of IID random variables approaches a normal distribution as the sample size becomes arbitrarily large. This is a very important result in statistical inference.
2. **Joint normality along with uncorrelatedness imply independence:** If X and Y are jointly normal and uncorrelated, then they are independent. This is because you can completely describe a normal distribution by its first two moments. If the covariance is zero, then the two variables are independent.

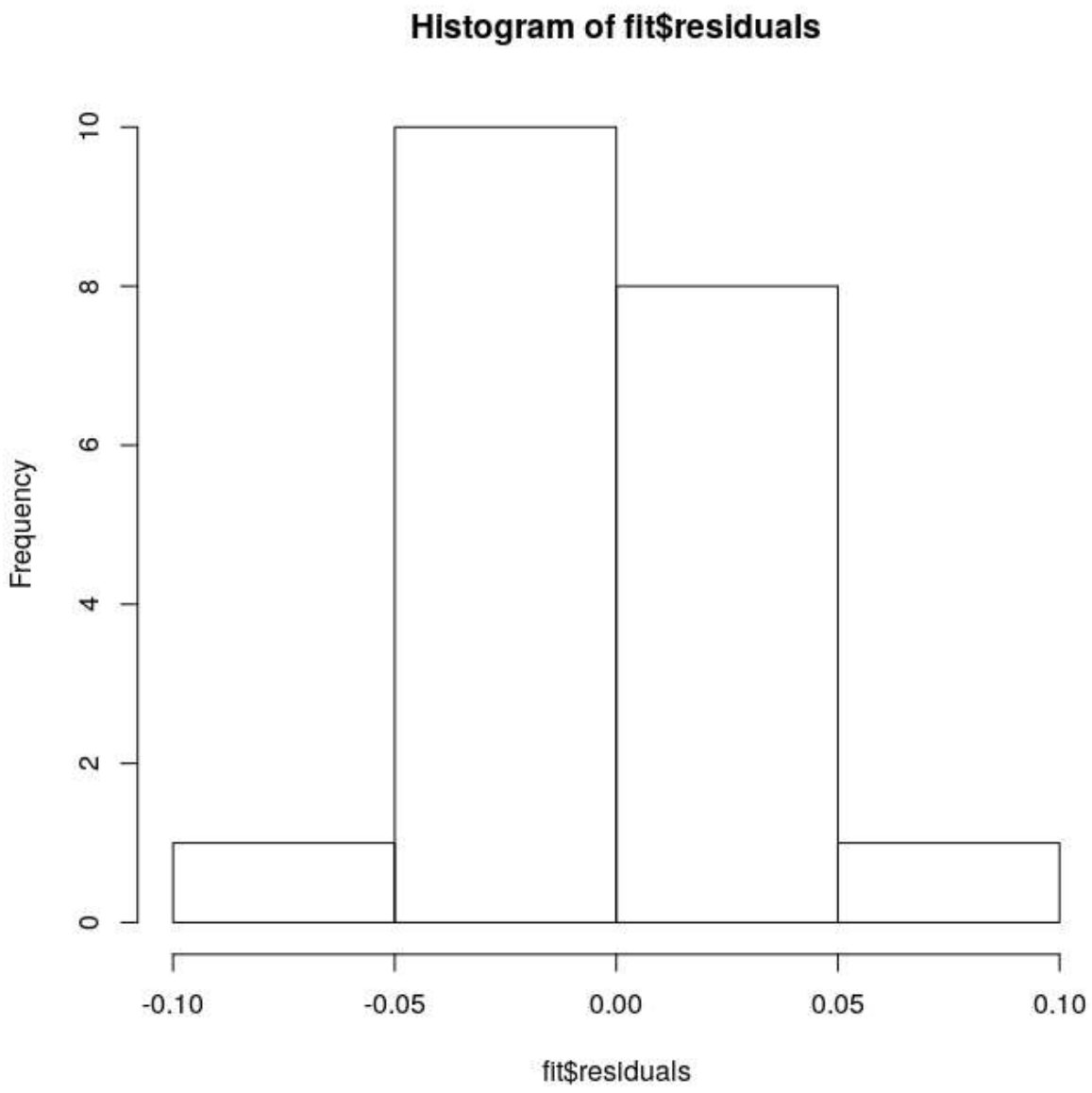
How to test for normality?

There are some rules to test for normality.

1. **Rule 1 of testing normality: Do not look at the Histogram and Kernel Density Plots:**

There isn't a more intricately complicated machinery ever discovered in the known universe than human brain. It should only be tasked with higher order abstract thinking such as making love, poetry, art, music, mathematics and photographing black holes. A complicated machine is not designed to look at granularity in efficient way. A theoretically normal distribution is flawlessly symmetrical and bell-shaped.

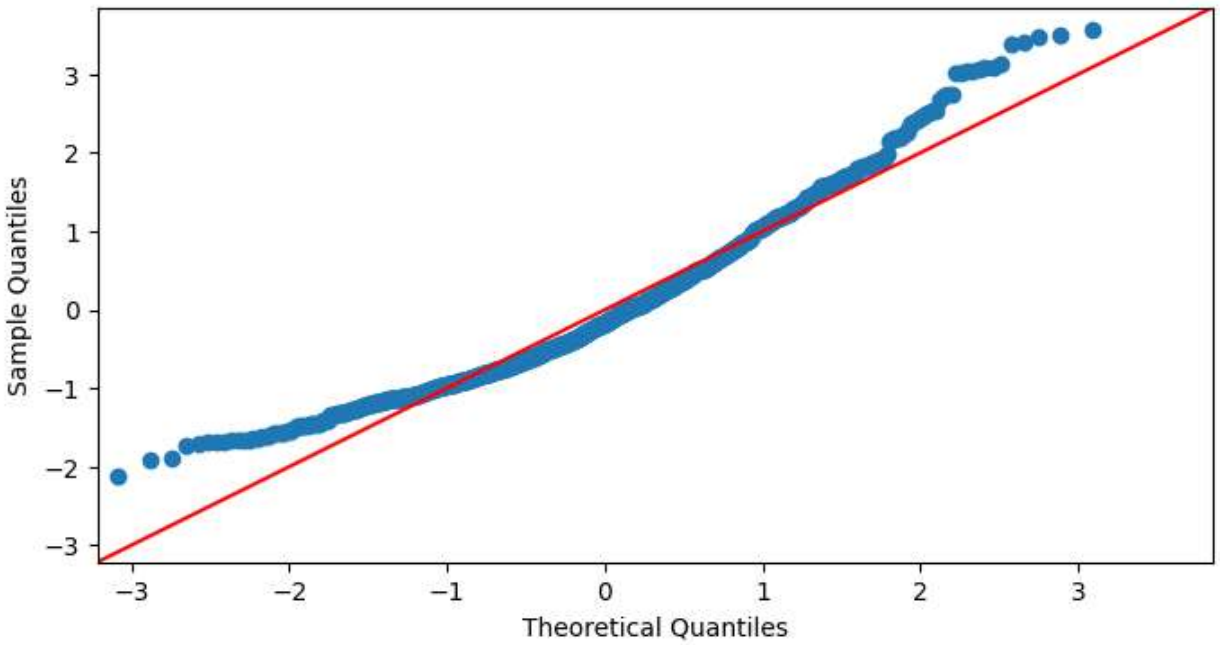
The histogram below looks normal to human eyes but it is far from normal.



With histograms and KDEs you need to worry about how many bins you are using, and how you are choosing the bin width. One of my professors once told me that your intuitions can trick you if you are not careful. So let us leave computational tasks to computers.

2. **Rule 2 of testing normality: Compare the quantiles of your data with theoretical normal quantiles:** A quantile represents the value below which a certain percentage of observations in a group of observations fall. For example, the 25th percentile is the value below which 25% of the observations may be found. A Q-Q plot shown below is a good way to compare the Quantiles. More you are off the line, more your data is not normal.

In [ ]:
```
from scipy.stats import skewnorm
import statsmodels.api as sm
import matplotlib.pyplot as plt
mean, std_dev, skewness = 0, 1, 4
samples = skewnorm.rvs(a=skewness, loc=mean, scale=std_dev, size=1000)
fig, ax = plt.subplots(figsize=(8, 4))
sm.qqplot(samples, fit=True, line='45', ax=ax)
plt.show()
```



3. **Rule 3 of Testing Normality: Rely on standard statistical tests but always read what is their null hypothesis and how they construct their test statistic:**

Two guys called Shapiro and Wilk, invented a procedure to test for normality called Shapiro-Wilk test. The null hypothesis is that the data is normally distributed. The test statistic is the W statistic. The test statistic is distributed as a chi-square random variable with $n-1$ degrees of freedom. The test statistic is calculated as follows:

$$W = \frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} x_i^2}$$

where x_i is the i^{th} observation and a_i is the i^{th} coefficient which is crafted in a way that it minimizes the variance of the test statistic given the null hypothesis is true.

Damage from a non-normal data:

The damage you incur is when you use a non-normal data to perform inferences that assume normality of the data. Although many inferences assume asymptotic normality and then bazaars are full of 'rule of thumbs'. I will leave this here because the damage is highly context-specific.

How to fix a non-normal data?

There is a easy way and a hard way.

The easy way: Use data transformations:

Transformations do not come for free. If you go for a transformation, ask the following questions:

1. Is the transformation possible? I once tried to calculate proportionate rate of change (growth rate) of temperature without realising many observations are negative and zeroes. Fun times!
2. Can you interpret the transformed data? If you apply a transformation, the definition of the variable changes. This can make interpretations hard.
3. How much information is lost with the transformation? If you apply a transformation, you are losing information. You need to be sure how much do you lose.

The hard way: Fix the source:

For this, I will be specific to finance. For many stocks, the periodic returns over a period of time are non-normal. One way to fix this is to change the definition of returns. Rather than calculating periodic returns, calculate the returns from either the volume bars or dollar bars. This means rather than doing a time sampling do a volume sampling or dollar sampling. For this, one must start with the tick data (data that records each transaction on an exchange) and then rather than sampling or aggregating at time, sample the prices when certain threshold number of shares are traded, which means if the threshold is 1000, then price series is created by sampling the price after every 1000 volume. This way of sampling usually exhibits desirable statistical propeerties. (Lopez de Prado, M. (2018). Advances in financial machine learning. John Wiley & Sons.). I really wanted to show this with real world data but then this write-up would become too long.

Stationarity

Technical Definition

Strong Stationarity is a property of a random process that the joint probability distribution does not change with time. This is a very strong assumption. A relaxed version of this assumption is of the weak stationarity. Now rather that talking about the entire distribution, we constrict

ourselves to the first two moments of a multivariate distribution. The first two moments are the vector of means and the co-variance matrix. Therefore, weak stationarity assumption is that:

1. The mean $\mu_t = \mu$ does not change with time.
2. The variance $\sigma_t^2 = \sigma^2$ does not change with time.
3. The autocovariance function $\gamma(t, t-k) = \gamma(k)$ only depends on the time difference and does not change with time.

Intuitive Definition

Stationarity means that statistical properties of a process do not change over time. Let us understand this step by step.

- A realisation of a time series (what we observe) falls out of a certain probability distribution which is usually unknown.
- The sequence of such realisations is called a stochastic process. Since we have a sequence of realisations, we need to talk about joint probability distribution of the realisations at different time instants.
- However, these realisations might not be independent of each other. For example, my mood right now is surely dependent on the mood I had one hour back. Moreover, the joint probability distribution could be a function of time itself. Overtime, I am learning to take things less seriously and more sincerely. Hence, my mood is becoming more and more stable.
- A fundamental part of predicting something is to understand the underlying probability distribution. If the probability distribution is changing with time, it is very difficult to predict the future.
- Therefore, we need to have that the joint probability distribution is not a function of time. This is the stationarity assumption.
- This assumption is very strong and is rarely satisfied in practice. We silly humans need to hammer this randomness down to fit the linear equations which are easier to work with. Therefore, we need to relax this assumption to a weaker version of stationarity and only restrict ourselves to the first two moments of the joint probability distribution.

Stationarity through an example

Let us talk about an AR(1) process which is the simplest class of linear stochastic processes. Imagine a drunkard walking. This drunkard has been hammered with some good Russian vodka. Now, the goal is to predict where this drunkard will land after taking certain number of steps, say 100 steps. We have a structure here! His each step is dependent on his previous step plus some random noise. The drunkard can only stretch his legs within some bounds and given the impact of comrade standard vodka, the next movement of his feet is independent of his previous movement. Hence, we can safely assume that this random noise is white noise that is it has no structure or is random, atleast in a linear sense. Formally, we can write this as:

$$X_{t+1} = X_t + e_{t+1}$$

X_t is the location of the drunkard at time t and e_{t+1} is the white noise with zero mean, σ^2 variance and zero covariance between succesive positions.

Each step of the drunkard adds some randomness to his trajectory. The marginal increase in his distance has a variance of σ^2. This variance keeps on cascading with each succesive step. The joint probability distribution here is a function of time.

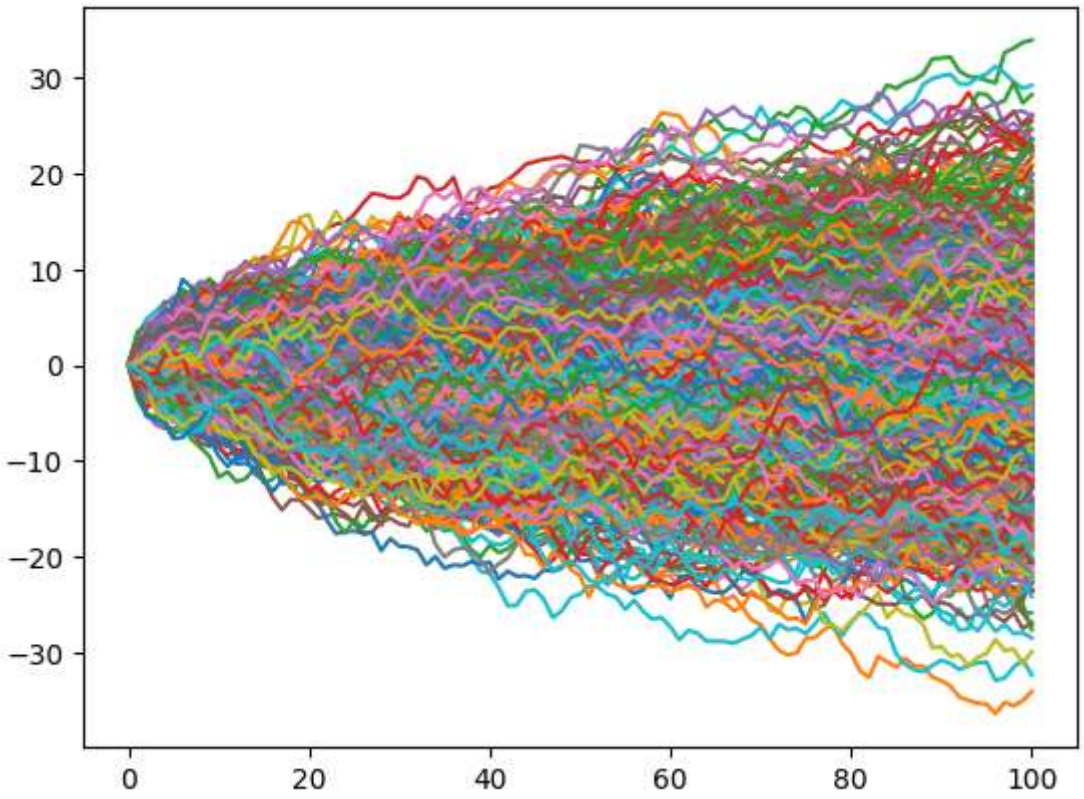
$$\begin{aligned}\text{Var}(Y_t) &= \text{Var}(e_1 + e_2 + \ldots + e_t) \\ &= \text{Var}(e_1) + \text{Var}(e_2) + \ldots + \text{Var}(e_t) \text{ (independence)} \\ &= \sigma^2 + \sigma^2 + \ldots + \sigma^2 = t\sigma^2\end{aligned}$$

Let us look at a simulatuion of this process. We will first define a function that will simulate the drunkard's walk. By the way, this is called a random walk. Maybe, the person who coined the term thought of the same example.

In [ ]:
```
import numpy as np
# Simulate a randomwalk
def generate_random_walk (steps: int, initial=0) -> list:
  walk = np.zeros(steps+1)
  walk[0] = initial
  for i in range(1, steps+1):
    walk[i] = walk[i-1] + np.random.normal(0,1)
  return walk
```

Let us simulate thousand of these drunkards, each taking 100 steps. You can see the plot below and appreciate how the variance increases with the number of steps.

In [ ]:
```
# Plot these 1000 random walks
import matplotlib.pyplot as plt
for _ in range(1000):
  plt.plot(generate_random_walk(100))
```



There are formal and not so formal ways to check for stationarity. The not-so-formal way is to just look at the time series plot and see if mean and variance is changing with time. The formal way is to do a statistical test. The most popular statistical test is the Augmented Dickey-Fuller test. The null hypothesis of this test is that the time series is non-stationary. If the p-value is less than 0.05, we reject the null hypothesis and conclude that the time series is stationary.

Dealing with non-stationarity

Random walk (also called unit-root processes) are a type of non-stationary process. Many wise women and men have claimed that the movement of many stock prices follow a random walk. For a random walk, the difference of successive realisations is a white noise which is stationary. We just discovered a way to make a process stationary. Differencing a time series is one of the most popular ways to make a non-stationary process stationary. The intuition is that if the process is non-stationary, the difference of successive realisations will be stationary. The transformation of stock prices into returns is an example of proportional differencing. The returns series is usually stationary, unless you are unlucky or are looking crypto-currencies.