New normality test based on the conditional second moments and the 20-60-20 rule

Damian Jelito

Based on the joint work with Marcin Pitera

damian.jelito@im.uj.edu.pl

Jagiellonian University in Kraków

27-th November, 2018

Outline

Several tests for normality

- Cramer-von Mises family

$$CvM := n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 w(x)\, \mathrm{d}F(x)$$

- Special case - Anderson-Darling

$$w_{AD}(x) := \frac{1}{F(x)(1 - F(x))}$$

- Computational version of the AD test statistic

$$AD := -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1) \left(\ln \left(\Phi \left(Y_{(i)} \right) \right) - \ln \left(1 - \Phi \left(Y_{(n-i+1)} \right) \right) \right)$$

with

$$Y_{(i)} := \frac{X_{(i)} - \bar{X}}{\hat{\sigma}}$$

Several tests for normality

- Shapiro-Wilk test

$$SW := \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}\left(X_{(i)} - \bar{X}\right)^2}$$

where

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}, \quad m_i = \mathbb{E}(X_{(i:n)}), \quad V_{ij} = \mathrm{Cov}(X_{(i:n)}, X_{(j:n)}).$$

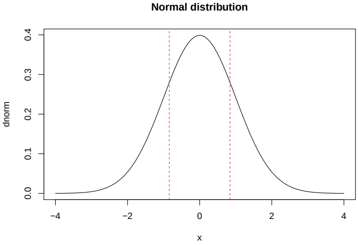
- Jarque-Bera test

$$JB := \frac{n}{6}\left(\hat{S}^2 + \frac{1}{4}(\hat{C} - 3)^2\right)$$

with

$$\hat{S} := \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{X}\right)^3}{\hat{\sigma}^3}, \quad \hat{C} := \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{X}\right)^4}{\hat{\sigma}^4}$$

The 20-60-20 Rule



$$\sigma_L^2 \quad = \quad \sigma_M^2 \quad = \quad \sigma_R^2$$

- We split the normal random variable into three disjoint conditioning sets: left (L), middle (M), and right (R):

$$L := \left(-\infty, F_X^{-1}(\tilde{q})\right],$$
$$M := \left(F_X^{-1}(\tilde{q}), F_X^{-1}(1-\tilde{q})\right),$$
$$R := \left[F_X^{-1}(1-\tilde{q}), +\infty\right)$$

- For a unique $\tilde{q} \approx 0.2$ the conditional variances coincide
- This property might be linked to the statistical phenomenon known as *The 20-60-20 Rule*.

Mathematical formulation of the 20-60-20 Rule

- Recall the conditional variance $\sigma_A^2 := \mathbb{E}((X - \mathbb{E}(X|A))^2|A)$

Theorem

If $X \sim \mathcal{N}(\mu, \sigma)$, then

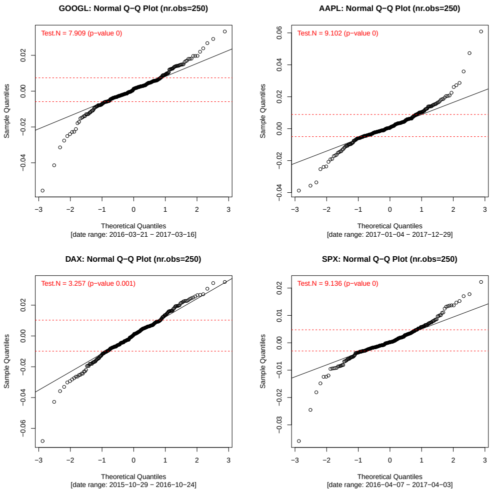
$$\sigma_L^2 = \sigma_M^2 = \sigma_R^2,$$

where $\tilde{q} = \Phi(x) \approx 0.19809$ and x is a unique negative solution of the equation

$$-x\Phi(x) - \phi(x)(1 - 2\Phi(x)) = 0.$$

- The property holds true for an arbitrary number of conditioning sets as well as in multivariate and elliptical case (for covariance matrices).

The 20-60-20 Rule and the Q-Q plot



1. We take return rates (based on adjusted daily close prices)
2. We make a simple Q-Q plot with theoretical normal distribution
3. We check if 20/60/20 division leads to accurate clustering

The test statistic

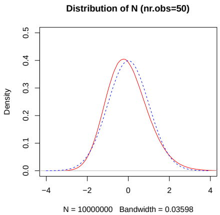
```
Test.N <- function(x){
    q1 <- quantile(x,0.2)
    q2 <- quantile(x,0.8)
    n  <- length(x)
    x.low  <- x[x <= q1]
    x.med  <- x[x > q1 & x < q2]
    x.high <- x[x >= q2]
    N <- var(x.low)+var(x.high)-2*var(x.med)
    N <- N * sqrt(n)/(var(x)*1.8)
    return(N)}
```

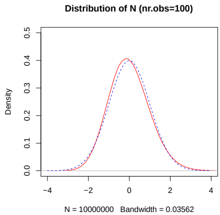
1. We introduce the test statistic

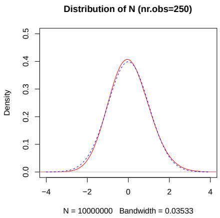
$$N := \frac{\sqrt{n}}{\rho}\left(\frac{\hat{\sigma}_L^2 - \hat{\sigma}_M^2}{\hat{\sigma}^2} + \frac{\hat{\sigma}_R^2 - \hat{\sigma}_M^2}{\hat{\sigma}^2}\right)$$

   where $\rho \approx 1.8$ is a fixed normalising constant
2. Under the normality assumption N is a pivotal quantity
3. N can be seen as a measure of tail fatness (cf. Anderson-Darling test)
4. N is based on the conditional second moments while Jarque-Bera test uses the third and fourth moments

Asymptotic distribution of the test statistic







α	n	$\Phi^{-1}(1-\alpha)$	$F_n^{-1}(1-\alpha)$
1.0%	50	2.33	2.64
	100		2.53
	250		2.47
2.5%	50	1.96	2.14
	100		2.08
	250		2.06
5.0%	50	1.64	1.74
	100		1.71
	250		1.71

Theorem

Let $X \sim N(\mu, \sigma)$. Then,

$$N \xrightarrow{d} \mathcal{N}(0,1), \qquad n \to \infty.$$

Moreover, ρ is independent of μ, σ and n.

We will come back to the asymptotics later.

Market data case study - overview

- We take S&P500 stocks returns from 01.2000 to 05.2018 (4610 daily adjusted close price returns for 381 stocks)
- For a given stock, the sample is split into disjoint sets of length n with $n \in \{50, 100, 250\}$.
- N is compared with Jarque–Bera test, Anderson–Darling test, and Shapiro–Wilk test.
- Normality hypothesis is checked at confidence level $\alpha \in \{1.0\%, 2.5\%, 5.0\%\}$.
- Non-normality of returns is a well known fact, hence we expect the null hypothesis to be rejected.
- We compute three supplementary metrics
  - **Statistic T** - **total rejection ratio** of a given test at confidence level α - for what proportion of all subsets the normality assumption was rejected.
  - **Statistic U** - **unique rejection ratio** of a given test at confidence level α - for what proportion of all subsets the normality assumption was rejected only by a given test (among all four tests).
  - **Statistic A** - **acceptance ratio** of a given test at confidence level α - for what proportion of all subsets the normality assumption was not rejected by any tests if it was not rejected by a given test.

Market data case study - results

Desc	nr runs	α	n	rejects	JB	AD	SW	N
T U A	35052	1.0%	50	31.5%	**25.9**% 1.9% **94.4**%	17.3% 0.6% 85.8%	23.2% 0.3% 91.7%	**25.9**% **3.1**% **94.4**%
T U A	35052	2.5%	50	39.6%	32.4% 2.4% **92.8**%	22.9% 0.9% 83.3%	28.3% 0.3% 88.7%	**32.5**% **3.9**% **92.8**%
T U A	35052	5.0%	50	47.9%	38.6% 2.4% 90.6%	29.1% 1.3% 81.1%	33.7% 0.4% 85.8%	**39.6**% **5.1**% **91.6**%
T U A	17526	1.0%	100	52.8%	45.2% 2.2% 92.5%	31.8% 0.6% 79.1%	41.3% 0.3% 88.6%	**46.1**% **4.4**% **93.4**%
T U A	17526	2.5%	100	61.3%	52.9% 2.2% 91.6%	38.8% 0.7% 77.5%	47.6% 0.2% 86.3%	**54.3**% **5.1**% **93.1**%
T U A	17526	5.0%	100	68.4%	59.7% 2.2% 91.3%	45.7% 0.8% 77.3%	53.4% 0.2% 84.9%	**61.3**% **5.3**% **92.9**%
T U A	6858	1.0%	250	88.5%	82.1% 1.0% 93.6%	71.2% 0.4% 82.7%	79.3% 0.1% 90.8%	**85.4**% **3.8**% **96.9**%
T U A	6858	2.5%	250	91.8%	86.8% 0.7% 95.1%	77.7% 0.2% 85.9%	83.7% 0.1% 91.9%	**89.4**% **3.0**% **97.6**%
T U A	6858	5.0%	250	93.9%	89.7% 0.5% 95.7%	82.4% 0.2% 88.5%	86.9% 0.0% 92.9%	**92.0**% **2.5**% **98.1**%

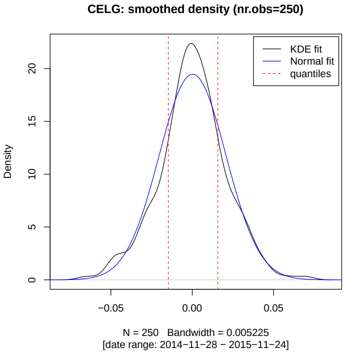
Metrics for the i-th test

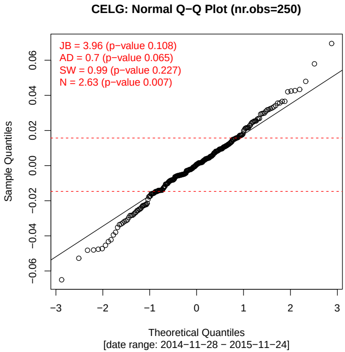
$$T_i = \frac{\#\{\text{i-th test rejected}\}}{\#\text{samples}},$$

$$U_i = \frac{\#\{\text{only i-th test rejected}\}}{\#\text{samples}},$$

$$A_i = \frac{\#\{\text{no test rejected}\}}{\#\{\text{i-th test didn't reject}\}}$$

Market data case study - close-up





Market data case study - p-value performance

Desc	nr runs	cond nr runs	n	JB	AD	SW	N
O	35052	16807	50	20.3%	7.8%	5.1%	28.7%
S				35.7%	17.2%	25.8%	—
O	17526	11994	100	15.0%	4.6%	3.2%	26.2%
S				26.1%	10.8%	18.7%	—
O	6858	6443	250	4.6%	1.4%	0.6%	15.4%
S				8.3%	3.4%	6.2%	—

$$O_i = \frac{\#\{\forall_j \text{ p.value}_i \leq \text{p.value}_j\}}{\#\text{samples}}$$

$$S_i = \frac{\#\{\text{p.value}_i \leq \text{p.value}_N\}}{\#\text{samples}}$$

- We compare p-values between different tests.
- For brevity, we consider only samples that were rejected by at least one test at level 5%.
- We present two performance measures:
  - **Statistic O - ratio of best p-values compared to other tests** - for what number of observations the p-value for a given statistic is smaller compared to all other p-values.
  - **Statistic S - ratio of best p-values compared to single test N** - for what number of observations the p-value for a given statistic is smaller compared to N test p-value.

Asymptotic distribution - notation

- $X \sim \mathcal{N}(\mu, \sigma)$ with mean parameter μ and standard deviation parameter σ
- $X_{(i)}$ - ith order statistic of the sample $(X_1, \ldots, X_n)$
- For $0 \leq \alpha < \beta \leq 1$ we define the conditioning set

$$A[\alpha, \beta] := \{x \in \mathbb{R} : F_X^{-1}(\alpha) < x \leq F_X^{-1}(\beta)\}.$$

- Recall $L := A[0, \tilde{q}]$, $R := A[1-\tilde{q}, 1]$, $M := A[\tilde{q}, 1-\tilde{q}]$,
- The conditional sample mean on set A

$$\overline{X}_A := \frac{1}{[n\beta] - [n\alpha]} \sum_{i=[n\alpha]+1}^{[n\beta]} X_{(i)}$$

- The conditional sample variance on set A

$$\hat{\sigma}_A^2 := \frac{1}{[n\beta] - [n\alpha]} \sum_{i=[n\alpha]+1}^{[n\beta]} \left(X_{(i)} - \overline{X}_A\right)^2$$

Main result

- Recall that the test statistic N is given by

$$N = \frac{1}{\rho}\left(\frac{\hat{\sigma}_L^2 - \hat{\sigma}_M^2}{\hat{\sigma}^2} + \frac{\hat{\sigma}_R^2 - \hat{\sigma}_M^2}{\hat{\sigma}^2}\right)\sqrt{n}.$$

- Restate our main result:

Theorem

Let $X \sim N(\mu, \sigma)$. Then,

$$N \xrightarrow{d} \mathcal{N}(0,1)\,, \qquad n \to \infty,$$

where ρ is a fixed normalising constant independent of μ, σ, and n.

- In the proof we make use of the series of lemmas.

Additional notation

For a fixed set $A = A[\alpha, \beta]$, we define

$$\begin{aligned}
\mu_A &:= \mathbb{E}[X | X \in A], \\
\sigma_A^2 &:= \mathbb{E}[(X - \mu_A)^2 | X \in A], \\
\kappa_A &:= \tfrac{1}{(\sigma_A^2)^2} \mathbb{E}[(X - \mu_A)^4 | X \in A], \\
a &:= F_X^{-1}(\alpha) = \mu + \sigma \Phi^{-1}(\alpha), \\
b &:= F_X^{-1}(\beta) = \mu + \sigma \Phi^{-1}(\beta).
\end{aligned}$$

Asymptotic normality of a conditional sample variance

Lemma (Asymptotic normality of a conditional sample variance)

For any $A = A[\alpha, \beta]$ it follows that

$$\sqrt{n}\left(\hat{\sigma}_A^2 - \sigma_A^2\right) \xrightarrow{d} \mathcal{N}(0, \tau_A),$$

where

$$\begin{aligned}\tau_A^2 := &\frac{1}{(\beta-\alpha)^2}\Bigg((\beta-\alpha)(\sigma_A^2)^2(\kappa_A - 1) + \alpha(1-\alpha)\left((a-\mu_A)^2 - \sigma_A^2\right)^2 \\ &- \alpha(1-\beta)\left((a-\mu_A)^2 - \sigma_A^2\right)\left((b-\mu_A)^2 - \sigma_A^2\right) \\ &+ \beta(1-\beta)\left((b-\mu_A)^2 - \sigma_A^2\right)^2\Bigg).\end{aligned}$$ [1]

[1] Note that for degenerate cases $\alpha = 0$ and $\beta = 1$, we get $a = -\infty$ and $b = \infty$, respectively. In those cases, the convention $0 \cdot \infty = 0$ should be used.

Additional lemmas

Lemma (Consistency of a conditional sample mean)
For any $A = A[\alpha, \beta]$ *it follows that* $\overline{X}_A \xrightarrow{\mathbb{P}} \mu_A, \quad n \to \infty.$

Lemma (Asymptotic normality of a conditional sample mean)
For any $A = A[\alpha, \beta]$ *it follows that*

$$\sqrt{n}\left(\overline{X}_A - \mu_A\right) \xrightarrow{d} \mathcal{N}(0, \eta_A), \quad n \to \infty,$$

where $0 < \eta_A < \infty$.

Some remarks on quantile estimators

Remark

We can replace $[n\beta] - [n\alpha]$ by $[n\beta] - [n\alpha] - 1$ in the definition of the conditional sample variance and our results remain valid.

Remark

Consider sequences (α_n) and (β_n) such that $n\alpha - \alpha_n$ and $\beta_n - n\beta$ are bounded. The corresponding conditional sample mean and variance is given by

$$\bar{X}_A^* := \frac{1}{\beta_n - \alpha_n} \sum_{i=\alpha_n+1}^{\beta_n} X_{(i)},$$

$$\hat{\sigma}_A^{2,*} := \frac{1}{\beta_n - \alpha_n} \sum_{i=\alpha_n+1}^{\beta_n} \left(X_{(i)} - \bar{X}_A^*\right)^2.$$

Then, we can replace $\overline{X}_A$ and $\hat{\sigma}_A^2$ by $\bar{X}_A^*$ and $\hat{\sigma}_A^{2,*}$ in our theorem and lemmas and their statements remain valid.

Some possible generalisations

- Different test statistic, e.g. $S_1 := \left(\frac{\hat{\sigma}_L^2 - \hat{\sigma}_R^2}{\hat{\sigma}_M^2}\right)\sqrt{n}$ or $S_2 := \left(\frac{\hat{\sigma}_L^2}{\hat{\sigma}_M^2} - \lambda\right)\sqrt{n}$
- More conditioning sets, e.g.

$$N_k := \sqrt{n}\left(\frac{\hat{\sigma}_{A_1}^2 - \hat{\sigma}_{A_k}^2}{\hat{\sigma}^2} + \ldots + \frac{\hat{\sigma}_{A_{k-1}}^2 - \hat{\sigma}_{A_k}^2}{\hat{\sigma}^2}\right)$$

  for partitioning sets $A_1, \ldots, A_k$ such that $\sigma_{A_1}^2 = \ldots = \sigma_{A_k}^2$ (exist for any $k \in \mathbb{N}$)
- Multivariate case, i.e. $\|\Sigma_A - \Sigma_B\|$, where $\|\cdot\|$ is a matrix norm and Σ_A is a conditional covariance matrix

Thank you for your attention!

References

[1] P. Jaworski, M. Pitera, *The 20-60-20 Rule*, Discrete Cont Dyn-B, Vol. 21 (2016), No 4, pp. 1149-1166.

[2] P. Jaworski, M. Pitera, *A note on conditional covariance matrices for elliptical distributions*, Stat Probabil Lett, Vol. 129C (2017), pp. 230-235.

[3] D. Jelito, M. Pitera, *New fat-tail normality test based on conditional second moments with applications to finance*, preprint, arXiv: 1811.05464

[4] H. Thode, *Testing for normality*, Marcel Dekker, 2002