1 Normality testing with the Wasserstein-2 distance

1.1 Classical normality tests

Normality or Goodnes of fit testing is a statistical procedure commonly used by practitioners such as lab/social scientists and economists. It's purpose is to determine whether it is reasonable to assume that an i.i.d. sample $X_1, ..., X_n$ comes from an unknown normal distribution. Once sufficient normality is established one is free to apply other procedures that operate under normality assumptions, such as the t-test to name one. Arguably the most commonly used such test is the Shapiro-Wilk test introduced in [?], which has been shown to have the greatest power in simulation studies [?] when compared to other common normality tests. The Shapiro-Wilk statistic W looks at the regression line of the order statistics $X_{(1)} \leq ... \leq X_{(n)}$ against the expected order statistics $m_1 \leq ... \leq m_n$ of a standard normal sample, i.e. the regression line through the normal q-q plot of the sample. Another, somewhat simpler test is the Shapiro-Francia test statistic which has the form

$$W' = \frac{\sum_{i=1}^{n}(X_{(i)} - \bar{X})(m_i - \bar{m})}{\sqrt{(\sum_{i=1}^{n}(X_{(i)} - \bar{X})^2)(\sum_{i=1}^{n}(m_{(i)} - \bar{m})^2)}} = \frac{\mathrm{Cov}\,(x, m)}{\sigma_x \sigma_m}.$$

1.2 Brownian bridge and Donsker's Theorem

Recall that a real-valued stochastic process $(B(t))_{t\in[0,1]}$ is called a Brownian bridge if it has the same distribution as $\{W(t)|W(1) = 0\}_{t\in[0,1]}$, where W is a Brownian motion. The Brownian bridge can be represented by

$$B(t) = W(t) - tW(1),$$

where W is a Brownian motion. Alternatively, it can be characterized as the unique centred Gaussian process with covariance function $\mathbb{E}\left[B(s)B(t)\right] = s \wedge t(1 - s \vee t)$. An improved version of Donsker's theorem states that if F is a cdf and F_n is its empirical version based on n i.i.d. samples then there exists a sequence of independent Brownian bridges $(B_n(t))_{n\geq 1}$ such that

$$\left\|\sqrt{n}(F_n - F) - B_n \circ F\right\|_\infty \preceq \frac{\log n}{\sqrt{n}}$$

almost surely.

1.3 Normality test based on W_2^2-distance

Suppose that we have an i.i.d. sample $X_1, ..., X_n \sim F$ and we want to test whether the X_i come from a normal distribution, i.e. whether $F \in \mathcal{H}_N$ where we define the normal location-scale family to be the set of distributions given by $\mathcal{H}_N := \{\Phi\left(\frac{\cdot - \mu}{\sigma}\right) : \mu \in \mathbb{R}, \sigma > 0\}$. In what follows I will write $\phi(\cdot)$ for the standard normal density, $\Phi(\cdot)$ for its cdf and F_n for the empirical cdf

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}.$$

Recall that if μ, ν are probability measures on $\mathbb{R}$ with cumulative distribution functions F_μ, F_ν respectively then the 2-Wasserstein distance between them can be written as

$$W_2^2(\mu, \nu) = \int_0^1 (F_\mu^{-1}(t) - F_\nu^{-1}(t))^2 dt,$$

where G^{-1} denotes the generalized inverse of the distribution function G defined as $G^{-1}(t) := \inf\{s : t \le F(s)\}$. Let us give a quick proof of this: first note that if $U \sim U[0,1]$ then $F_\mu^{-1}(U) \sim \mu$, $F_\nu^{-1}(U) \sim \nu$ and their joint distribution function is given by $\mathbb{P}\left(F_\mu^{-1}(U) \le x, F_\nu^{-1}(U) \le y\right) = F_\mu(x) \wedge F_\nu(y)$. Let $\Gamma(\mu, \nu)$ denote the set of couplings of (μ, ν). Now, by definition of the Wasserstein distance we have

$$W_2^2(\mu, \nu) = \inf_{(X,Y) \in \Gamma(\mu,\nu)} \mathbb{E}\left[(X - Y)^2\right] = \mathbb{E}\left[X^2\right] + \mathbb{E}\left[Y^2\right] - 2 \sup_{(X,Y) \in \Gamma(\mu,\nu)} \mathbb{E}\left[XY\right],$$

so that the above is minimized when $\mathbb{E}\left[XY\right]$ is maximized. Recall that for non-negative random varibles X, Y it holds that

$$\mathbb{E}\left[XY\right] = \int_0^\infty \int_0^\infty \mathbb{P}\left(X > x, Y > y\right) dx dy.$$

The final observation to make is that

$$\mathbb{P}\left(X > x, Y > y\right) = \mathbb{P}\left(X > x\right) + \mathbb{P}\left(Y > y\right) - 1 + \mathbb{P}\left(X \le x, Y \le y\right),$$

so that $\mathbb{P}\left(X > x, Y > y\right)$ is maximal iff $\mathbb{P}\left(X \le x, Y \le y\right)$ maximal. But trivially $\mathbb{P}\left(X \le x, Y \le y\right) \le F_\mu(x) \wedge F_\nu(y)$ so that we see the coupling $(F_\mu^{-1}(U), F_\nu^{-1}(U))$ must be optimal. The extension to negative and unbounded random variables is straightforward.

Let F be the distribution function of some square-integrable probability measure $P \in \mathcal{P}_2$ and consider its distance to the normal location-scale family w.r.t. the 2-Wasserstein distance. Assuming that the mean of P is μ_0 and its variance is σ_0^2

we have

$$\begin{aligned} W_2^2(F,\mathcal{H}_N) &:= \inf\{W_2^2(P,H)\,:\,H\in\mathcal{H}_N\} \\ &= \inf_{\mu\in\mathbb{R},\,\sigma>0}\int_0^1 (F^{-1}(t)-\mu-\sigma\Phi^{-1}(t))dt \\ &= \inf_{\mu\in\mathbb{R},\,\sigma>0} \sigma_0^2+(\mu_0-\mu)^2+\sigma^2-2\sigma\int_0^1 F^{-1}(t)\Phi^{-1}(t)dt \\ &= \sigma_0^2-\left(\int_0^1 F^{-1}(t)\Phi^{-1}(t)dt\right)^2, \end{aligned}$$

where the minimizing values are $\mu=\mu_0$ and $\sigma=\int_0^1 F^{-1}(t)\Phi^{-1}(t)dt$. Notice that the quantity $\frac{W_2^2(F,\mathcal{H}_N)}{\sigma_0^2}$ is unaffected by location and scale changes in P.

Suppose now that F is the distribution of our sample and let us replace it by the empirical cdf F_n. This gives us our goodness-of-fit test statistic

$$\mathcal{R}_n = 1-\frac{\left(\int_0^1 F_n^{-1}(t)\Phi^{-1}(t)dt\right)^2}{S_n^2} =: 1-\frac{\hat{\sigma}_n^2}{S_n^2},$$

where S_n^2 is the usual sample variance. We want to analyze the asymptotic distribution of this random variable under the null hypothesis that $P\in\mathcal{H}_N$. By location-scale invariance we can assume without loss of generality that $P=\Phi$. In fact, instead of $\mathcal{R}_n$ it will be conventient to look at

$$\mathcal{R}_n^* := S_n^2\mathcal{R}_n.$$

However, since $P=\Phi$ we know that $S_n^2\overset{a.s.}{\to}1$ so by Slutsky's lemma we can analyze either of $\mathcal{R}_n^*$ and $\mathcal{R}_n$ and arrive at the same asymptotic properties. The following is the main result.

Theorem 1.1 — *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. standard normal random variables. Then for a Brownian bridge $(B(t))_{t\in[0,1]}$ we have*

$$\begin{aligned} n(\mathcal{R}_n^*-a_n) \overset{d}{\to} &\int_0^1 \frac{B(t)^2-\mathbb{E}\left[B(t)^2\right]}{\phi^2(\Phi^{-1}(t))}dt-\left(\int_0^1\frac{B(t)}{\phi(\Phi^{-1}(t))}dt\right)^2 \\ &-\left(\int_0^1\frac{B(t)\Phi^{-1}(t)}{\phi(\Phi^{-1}(t))}dt\right)^2, \end{aligned}$$

where a_n is given by

$$a_n=\frac{1}{n}\mathbb{E}\left[\int_{I_n}\frac{B(t)^2}{\phi^2(\Phi^{-1}(t))}dt\right].$$

Furthermore, the above is equal in distribution to

$$-\frac{3}{2}+\sum_{j=3}^{\infty}\frac{Z_j^2-1}{j}, \tag{1.1}$$

where the Z_i are independent standard normal random variables.

Remark 1.1. What's surprising about the above is that in [**?**] it was shown that the Shapiro-Wilk and Shapiro-Francia test have the same asymptotic distribution.

Remark 1.2. At a first glance it might not even seem obvious that expression (1.1) is well defined. However, looking at the second moment of the partial sums it can be verified that the partial sums are bounded in L^2 and thus uniformly integrable. Taking expectation now gives a value of $-3/2$ so that in particular the random variable is finite almost surely.

1.4 Sketch proof of Theorem 1.1

The following decomposition is essential for the proof:

$$n\mathcal{R}_n^* = \int_0^1 \left(\frac{\rho_n(t)}{\phi(\Phi^{-1}(t))}\right)^2 dt - \left(\int_0^1 \frac{\rho_n(t)}{\phi(\Phi^{-1}(t))} dt\right)^2 \tag{1.2}$$

$$- \left(\int_0^1 \frac{\rho_n(t)\Phi^{-1}(t)}{\phi(\Phi^{-1}(t))} dt\right)^2 \tag{1.3}$$

$$=: \mathcal{R}^{(1)} - \mathcal{R}^{(2)} - \mathcal{R}^{(3)}, \tag{1.4}$$

where $\rho_n(t) := \sqrt{n}\phi(\Phi^{-1}(t))(\Phi^{-1}(t) - F_n^{-1}(t))$ is a weighted quantile process. The main ingredient of the proof is the following theorem, which is reminiscent of Donsker type theorems:

Proposition 1.1 — *There exist a sequence of Brownian bridges $\{B_n(t),\, t \in [0,1]\}_{n\geq 0}$ such that for any $\nu \in (0, 1/2]$,*

$$\sup_{t\in I_n} \frac{|\rho_n(t) - B_n(t)|}{(t(1-t))^\nu} \leq \mathcal{O}_P(1) n^{\nu - 1/2},$$

where I_n is the interval $(\frac{1}{n+1}, \frac{n}{n+1})$.

The following lemma is the last ingredient of the proof:

Lemma 1.2 — *Suppose $\{X_n\}_{n\geq 1}$ is an i.i.d. sample from a standard normal distribution. Then*

$$n\int_0^{1/(n+1)} (X_{(1)} - \Phi^{-1}(t))^2 dt = o_P(1) \qquad \text{and} \qquad n\int_{n/(n+1)}^1 (X_{(n)} - \Phi^{-1}(t))^2 dt = o_P(1),$$

as $n \to \infty$.

Sketch proof of Theorem 1.1. Notice that Lemma 1.2 allows us to change the range over which we integrate in 1.3 from $[0, 1]$ to I_n and only incur a penalty of $o_P(1)$

and we get

$$
\begin{aligned}
n\mathcal{R}_n^* - \int_{I_n} \left(\frac{\rho_n(t)}{\phi(\Phi^{-1}(t))} \right)^2 dt + \left(\int_{I_n} \frac{\rho_n(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 \\
+ \left(\int_{I_n} \frac{\rho_n(t)\Phi^{-1}(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 = o_P(1).
\end{aligned}
$$

Thus, if we can show that

$$
\begin{aligned}
A_1 &:= \int_{I_n} \left(\frac{\rho_n(t)}{\phi(\Phi^{-1}(t))} \right)^2 dt - \int_{I_n} \left(\frac{B_n(t)}{\phi(\Phi^{-1}(t))} \right)^2 dt = o_P(1) \\
A_2 &:= \left(\int_{I_n} \frac{\rho_n(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 - \left(\int_{I_n} \frac{B_n(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 = o_P(1) \\
A_3 &:= \left(\int_{I_n} \frac{\rho_n(t)\Phi^{-1}(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 - \left(\int_{I_n} \frac{B_n(t)\Phi^{-1}(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 = o_P(1)
\end{aligned}
$$

we are done. Notice that the only difference between the integrals are that ρ_n is replaced by B_n. Obviously, this is where Proposition 1.1 comes in handy. □