MS = GOODNESS-OF-FIT TESTING

I - EMPIRICAL DISTRIBUTION FUNCTION

Let $X_1, \dots, X_n$ be an iid sample, where each $X_j \sim P$, with distribution function F. Let $X_{(1)}, \dots, X_{(n)}$ be the ordered sample $X_{(1)} \leq \dots \leq X_{(n)}$.

Note that if P is AC with density f, then $X_{(1)}, \dots, X_{(n)}$ is a sufficient statistic for F. Indeed, having observed $S(X_1, \dots, X_n) := (X_{(1)}, \dots, X_{(n)})$, the only possible values for the original sample are the $n!$ different permutations of $(X_{(1)}, \dots, X_{(n)})$. Each of these permutations have a probability $1/n!$, which is independent of the parameter(s) of the distribution P/f.

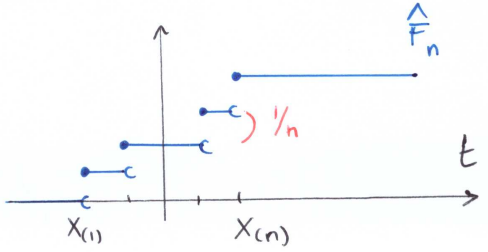
→ We construct an estimator of F based on the order statistics $X_{(1)}, \dots, X_{(n)}$. A natural candidate is the EMPIRICAL DISTRIBUTION FUNCTION (EDF):

$$\hat{F}_n(t) := \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}(X_j \leq t) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}(X_{(j)} \leq t), \quad t \in \mathbb{R},$$

which assigns a probability $1/n$ to each observation X_j.

(The associated empirical distribution is denoted $\hat{P}_n = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j}$)

↑ Dirac at X_j.



The SLLN gives us that $\forall t \in \mathbb{R}$,

$$\hat{F}_n(t) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}(X_j \leq t) \xrightarrow[n \to +\infty]{a.s.} \mathbb{E}\, \mathbb{1}(X \leq t) = F(t).$$

↑ sum of n iid RVs

The convergence is in fact uniform over t:

Theorem (GLIVENKO-CANTELLI)

$$\sup_t \left| \hat{F}_n(t) - F(t) \right| \xrightarrow[n \to +\infty]{a.s.} 0$$

↘ So we can write $\hat{F}_n(t) = F(t) + o(1)$, where the term "$o(1)$" does not depend on t.

proof: There exists several approaches to prove the GC theorem. One possible way is to show the existence of positive constants c_1, c_2 such that $\forall \varepsilon > 0$,

$$(*) \quad \mathbb{P}\left(\sup_{t \in \mathbb{R}} \left| \hat{F}_n(t) - F(t) \right| > \varepsilon \right) \leq c_1 n e^{-c_2 n \varepsilon^2},$$

following the lines in the proof of the VC inequality. (see chapter SL: VAPNIK CHERVONENKIS THEORY)

Almost sure convergence follows from Borel-Cantelli lemma.

For more information on how to get (*), we refer to Theorem 12.4 in Devroye, Györfi & Lugosi (1996), A Probabilistic Theory of Pattern Recognition. Springer. ∎

The RVs $\mathbb{1}(X_j \leq t)$ are $B(F(t))$ distributed, with mean $F(t)$ and variance $F(t)(1-F(t))$. A CLT holds for the EDF:

$$n^{1/2}\left(\hat{F}_n(t) - F(t) \right) \xrightarrow[n \to +\infty]{d} \mathcal{N}(0,\ F(t)(1-F(t))),$$

$\forall t \in \mathbb{R}$.

× Observation:

$$\hat{F}_n(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(X_j \le t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(U_j \le F(t)) =: \hat{U}_n(F(t)),$$

$X_j \sim F$
$\Rightarrow X_j = F^-(U_j)$, $U_j \sim \mathcal{U}(0,1)$
where F^- = generalized inverse of F.

where $\hat{U}_n(u)$, $u \in [0,1]$, is the EDF of a uniform $(0,1)$ random sample $U_1, \dots, U_n$.

⇒ Properties of $\hat{F}_n$ can be deduced from that of $\hat{U}_n$.
In particular, $\forall u \in [0,1]$,

$$n^{1/2}\left(\hat{U}_n(u) - u\right) \xrightarrow{d} V(u) \sim \mathcal{N}(0, u(1-u)),$$

and a multivariate version holds as well: for any $0 \le u_1 < \dots < u_d \le 1$,

$$n^{1/2}\left(\hat{U}_n(u_1) - u_1, \dots, \hat{U}_n(u_d) - u_d\right) \xrightarrow{d} \left(V(u_1), \dots, V(u_d)\right) \overset{2}{\sim} MN(0, \Sigma(u)),$$

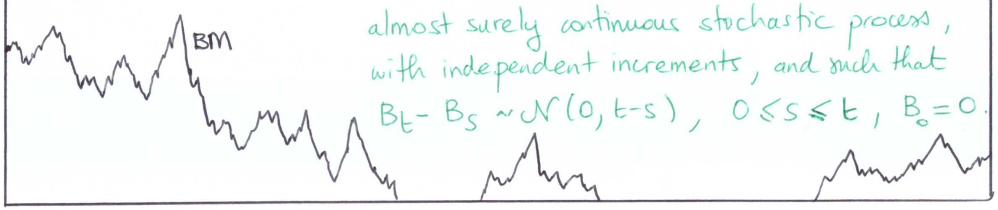
where $\Sigma(u) = \left(\min(u_j, u_k)\left(1 - \max(u_j, u_k)\right)\right)_{j,k=1,\dots,d}$.

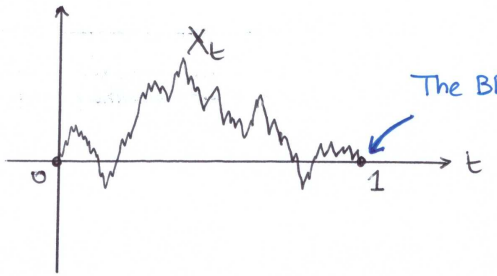
Same covariance matrix as for the BROWNIAN BRIDGE (BB).

× Interlude: Background on Brownian Bridge (BB).

A BB is a conditional Brownian Motion (BM); aka a "tied-down" BM, and is defined as $X_t \overset{d}{=} (B_t \mid B_1 = 0)$, for $0 \le t \le 1$, where B_t is a BM.

aka the WIENER PROCESS: an almost surely continuous stochastic process, with independent increments, and such that $B_t - B_s \sim \mathcal{N}(0, t-s)$, $0 \le s \le t$, $B_0 = 0$.





The BB is tied-down at $t=0$ and $t=1$.

BB belongs to the class of diffusion processes. It satisfies the STOCHASTIC DIFFERENTIAL EQUATION (SDE):

$$dX_t = -\frac{X_t}{1-t}dt + dB_t, \quad X_0 = 0, \quad 0 \le t \le 1$$

loosely speaking, this expression means that for small $h > 0$,

$$X_{t+h} - X_t \mid X_t = x \approx \mathcal{N}\left(-\frac{xh}{1-t}, h\right)$$

As t gets closer to 1, there is a stronger and stronger force pulling the process back towards 0.

A BB is a Gaussian Process (Markovian + normal transition densities) ⇒ it is completely determined by its mean function $m_X(t) = \mathbb{E}X_t$, and covariance function $r_X(s,t) = \mathrm{Cov}(X_s, X_t)$.

$$\searrow X_t = X_0 - \int_0^t \frac{X_s\,ds}{1-s} + B_t \quad \text{Taking } \mathbb{E}(\cdot)$$

$$m_X(t) = -\int_0^t \frac{m_X(s)}{1-s}ds \quad \text{Differentiating}$$

$$m_X'(t) = -\frac{m_X(t)}{1-t} \quad \text{& solving} \quad \frac{m_X'(t)}{m_X(t)} = \frac{d\log m_X(t)}{dt} = -\frac{1}{1-t}$$

$$\Rightarrow \log m_X(t) = \log(1-t) + C \Rightarrow m_X(t) = C(1-t).$$

Since $m_X(0) = 0 = C(1-0) \Rightarrow C = 0 \Rightarrow \boxed{m_X(t) = 0}$

So the BB is a zero mean Gaussian process.

↳ To get the covariance function, one proceed similarly, noticing that $dX_t^2 = \left(1 - \frac{2X_t^2}{1-t}\right)dt + 2X_t\,dB_t$, by Itô's formula. Taking expectations & differentiating, yields $\mathrm{Var}\,X_t = t(1-t)$. Finally, for $0 \le s \le t \le 1$, considering $X_s X_t = X_s\left(X_s + \int_s^t dX_u\right)$, taking $\mathbb{E}(.)$ and differentiating $\frac{\partial}{\partial t}$ keeping s fixed yields a differential equation for $r_X(s,t)$, whose solution is $r_X(s,t) = C_s(1-t)$. Since $r_X(s,s) = s(1-s)$, we get $C_s = s$, and

$$\boxed{r_X(s,t) = s(1-t), \quad 0 \le s \le t \le 1}$$

From the mean & covariance function, we deduce an alternative definition of BB: a Gaussian process $\{X_t\}$ with $\mathbb{E}X_t = 0$, and $\mathbb{E}(X_s X_t) = s(1-t)$, $0 \le s \le t \le 1$.

↖ which also corresponds to the mean & covariance function of $B_t - tB_1$, $t \in [0,1]$, and of $(1-t)B_{t/(1-t)}$.

Thus

$$X_t \text{ is a BB} \iff X_t = B_t - tB_1 \iff X_t = (1-t)B_{t/1-t}.$$

End of Interlude ■

Recall that for the EDF $\hat{U}_n$, holds that

$$n^{1/2}\left(\hat{U}_n(s) - s,\ \hat{U}_n(t) - t\right) \xrightarrow{d} \left(V(s), V(t)\right) \quad 0 \le s \le t \le 1,$$

where $(V(s), V(t))$ is MN with cov matrix $\begin{pmatrix} s(1-s) & s(1-t) \\ s(1-t) & t(1-t) \end{pmatrix}$.

This actually suggests that if we consider $\{n^{1/2}(\hat{U}_n(u) - u)\}_{u \in [0,1]}$ as a random process on $[0,1]$, then it converges in distribution in the space $C[0,1]$ of continuous functions on $[0,1]$ to the Brownian Bridge (BB) process $\{V(u)\}_{u \in [0,1]}$.

↖ Statement requires a proof !

In particular,

Theorem = $\sup_{u \in [0,1]} |n^{1/2}(\hat{U}_n(u) - u)| \xrightarrow{d} \max_{u \in [0,1]} |V(u)|$,

where $\{V(u)\}_{u \in [0,1]}$ is a BB.

And the distribution of the RHS is known:

$$\mathbb{P}\left(\max_{u \in [0,1]} |V(u)| \le x\right) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2}, \quad x > 0$$

denoted $K(x) :=$ KOLMOGOROV FUNCTION

Corollary : (KOLMOGOROV+) / (DONSKER) since $\sup_t |\hat{F}_n(t) - F(t)| \overset{d}{=} \sup_{u \in [0,1]} |\hat{U}_n(u) - u|$

$$n^{1/2} \sup_t |\hat{F}_n(t) - F(t)| \xrightarrow{d} \max_{u \in [0,1]} |V(u)|,$$

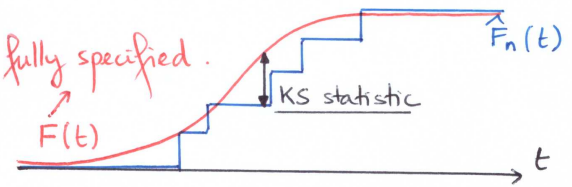
with $\mathbb{P}\left(\max_{u \in [0,1]} |V(u)| \le x\right) = K(x) =$ Kolmogorov function, $x > 0$.

↖ We discuss next implications of this result for testing goodness of fit.

II. TESTS ON 1 SAMPLE

II.1. EDF tests

The general principle of EDF tests is to compare a theoretical distribution function $F(t)$ specified under a null hypothesis H_0 with the EDF $\hat{F}_n(t)$.



The KOLMOGOROV-SMIRNOV (KS) statistic is based on the maximum difference between $F(t)$ and $\hat{F}_n(t)$:

$$D_n = \sup_t |\hat{F}_n(t) - F(t)|$$

And the good news is that the distribution of $n^{1/2} D_n$ for n large is approximately known, and follows from the Kolmogorov + Donsker theorem on page 6

$\Rightarrow$ Reject H_0 at level α if $n^{1/2} D_n > K_{1-\alpha}$, where K_α is the α-quantile of the Kolmogorov distribution function

$$K(x) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2x^2}, \quad x > 0.$$

A distribution that does not depend on F.

For small values of n, tables exist.

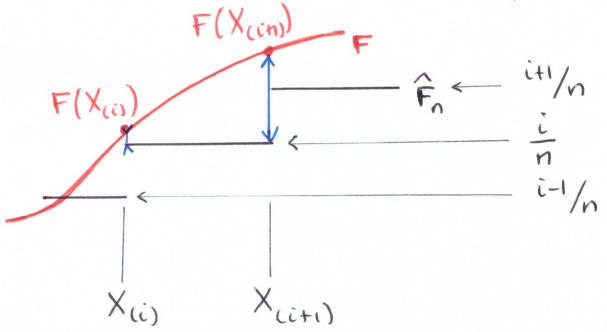
- Alternatively, you may consider the statistic $D_n^+ := \sup_t (\hat{F}_n(t) - F(t))$ (no absolute values)

You are then testing for $H_0: X_i \sim F$ versus the alternative $H_1: X_i \sim F_1 \geq F$. The asymptotic distribution of $n^{1/2} D_n^+$ is known as well $\mathbb{P}(n^{1/2} D_n^+ < x) \xrightarrow{n \to +\infty} 1 - e^{-2x^2}$ (Smirnov '42)

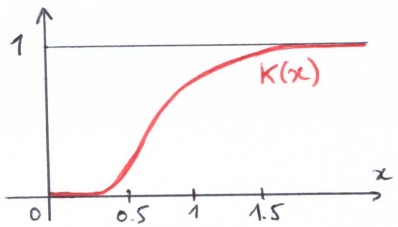
Remarks: (i) The value $D_n = \sup_t |\hat{F}_n(t) - F(t)|$ is computed in practice using the simpler expression:

$$D_n = \max_{0 \leq i \leq n} \max \left\{ \left| \frac{i}{n} - F(X_{(i)}) \right| ; \left| \frac{i}{n} - F(X_{(i+1)}) \right| \right\}.$$

where $X_{(1)} \leq \dots \leq X_{(n)}$, and $X_{(0)} := -\infty$; $X_{(n+1)} := +\infty$.



(ii) Remark on Kolmogorov function: the sum $\sum_{k=1}^{\infty} (-1)^k e^{-2k^2x^2}$ is slow to converge; $\approx$ a hundred terms are needed to obtain a good numerical approximation.



x	K(x)
0.5	0.036
0.8	0.456
1	0.730
1.5	0.978
1.8	0.996
2	0.999

Note that the null distribution F is completely specified (e.g. $\mathcal{N}(0,1)$, $\mathcal{P}(1)$, $\mathcal{B}(1/2)$, ...); i.e. there are no unknown parameters.

→ What if we want to test "Does X have a Gaussian distribution?", but we don't know the parameters?
A simple idea is to estimate the unknown parameters, and to plug them back-in.

Ex: $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

(*) Consider $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ & $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu})^2$, and compare $\hat{F}_n(t)$ with the distribution function of a $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ RV: $\sup_t \left| \hat{F}_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t) \right|$.

⚠ Kolmogorov + Donsker Theorem is no longer valid!
A common mistake!
⇒ The distribution of $\sup_t |\hat{F}_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t)|$ is not (approximately) given by Kolmogorov function.

→ Instead, use the KOLMOGOROV-LILLIEFORS test: proceed as in (*) above, and estimate the quantiles of the distribution of $\sup_t |\hat{F}_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t)| =: \hat{D}_n$ using Monte-Carlo simulations:
- ↳ Under H_0, the true distribution is $\Phi_{\hat{\mu},\hat{\sigma}^2}$.
- ↳ Generate B samples $\sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ of size n
- ↳ Compute the quantities $\sup_t |\hat{F}_n^b(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t)| =: Y_b$ where $\hat{F}_n^b$ = EDF of the b-th sample, $b = 1, \dots, B$.
- ↳ Estimate the distribution of $\hat{D}_n$ from $Y_1, \dots, Y_B$.

• Another commonly used goodness of fit test based on the EDF is the VON-MISES-SMIRNOV test, based on the statistic

$$\omega_n^2 := n \int \left(\hat{F}_n(t) - F(t) \right)^2 dF(t)$$

↑ $\mathbb{E}(\ell_2 \text{ error})$

↖ Compared to the KS test, which relies only on the largest value of the discrepancy between $\hat{F}_n$ and F, the error $\forall t$ is integrated out.

Since $\omega_n^2 \overset{d}{=} \int_0^1 \left(n^{1/2} [\hat{U}_n(u) - u] \right)^2 du$, convergence of the empirical process $\{n^{1/2}(\hat{U}_n(u) - u)\}_{u \in [0,1]}$ to the BB implies that $\mathbb{P}(\omega_n^2 \le x) \xrightarrow{n \to +\infty} \mathbb{P}\left(\int_0^1 |V(u)|^2 du \le x \right)$, which is known / tabulated.

→ In practice, the stat. ω_n^2 is calculated using the simpler expression $\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^n \left(\frac{2i-1}{2n} - F(X_{(i)}) \right)^2$, where $X_{(1)} \le \dots \le X_{(n)}$.

• Alternatively, one uses the ANDERSON-DARLING statistic

$$A = n \int \frac{\left(\hat{F}_n(t) - F(t) \right)^2}{F(t)(1 - F(t))} dF(t)$$

↖ Renormalized by the variance of the empirical process.

, whose distribution is also tabulated.

Note that $A = -n - \frac{1}{n}\sum_{i=1}^n \{(2i-1)\log F(X_{(i)}) + (2n+1-2i)\log(1 - F(X_{(i)}))\}$.

II.2. Shapiro-Wilk test

The Shapiro-Wilk test is a test for normality based on L-STATISTICS.

↳ a statistic that is a Linear combination of order statistics

Ex: median, mean, max, quantile ...

⊕ Robust to outliers, easy to compute

⊖ Usually less efficient than other estimators

Ex: Estimating the mean μ of a normal population. The most efficient estimator is the sample mean $\bar{X}$. For a symmetric distribution, the median is equal to the mean, and so we can take

$$\tilde{X} := \begin{cases} X_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(X_{(n/2)} + X_{1+(n/2)}\right) & \text{if } n \text{ is even} \end{cases}$$

as an estimator of μ. Comparing the variances of $\bar{X}$ and $\tilde{X}$, if $n = 2k+1$, then it is possible to show that $\operatorname{Var}\bar{X} / \operatorname{Var}\tilde{X} = \frac{4k}{\pi(2k+1)} \to \frac{2}{\pi} \approx 0.67$, as $k \to +\infty$.

The Shapiro-Wilk test compares the empirical variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ with an estimator of the variance having known properties under the assumption that the $X_1, \dots, X_n$ are normally distributed.

Let $X_1, \dots, X_n$ iid with mean μ & variance σ^2.

$X_{(1)} \leq \dots \leq X_{(n)}$ = ordered sample

Put $Z_i = \frac{X_i - \mu}{\sigma} \Rightarrow Z_{(i)} = \frac{X_{(i)} - \mu}{\sigma}$

Let $\alpha_i := \mathbb{E} Z_{(i)}$ (mean & covariance of ordered $\mathcal{N}(0,1)$ RVs)

$B_{ij} := \operatorname{Cov}\left(Z_{(i)}, Z_{(j)}\right)$

Then $X_{(i)} = \mu + \sigma Z_{(i)}$

$= \mu + \sigma \alpha_i + \varepsilon_i$, writing $Z_{(i)} = \mathbb{E} Z_{(i)} + \frac{\varepsilon_i}{\sigma}$

where the ε_i are zero mean & correlated RVs, with $\operatorname{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 B_{ij} = \mathbb{E}(\varepsilon_i \varepsilon_j)$.

Putting $1 = \underset{(n\times 1)}{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}$, $\alpha = \underset{(n\times 1)}{\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}}$, $A = \underset{(n\times 2)}{\begin{pmatrix} 1 & \alpha_1 \\ \vdots & \vdots \\ 1 & \alpha_n \end{pmatrix}}$, $X_{(\cdot)} = \underset{(n\times 1)}{\begin{pmatrix} X_{(1)} \\ \vdots \\ X_{(n)} \end{pmatrix}}$

and $\varepsilon = \underset{(n\times 1)}{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}$, we have $\boxed{X_{(\cdot)} = A\begin{pmatrix} \mu \\ \sigma \end{pmatrix} + \varepsilon}$ ← residual error

"Response variable" "predictors" vector of coefficients to estimate

⇒ Estimate (μ, σ^2) using weighted least squares

since the errors are correlated, with varying variances

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} := \underset{\mu,\sigma}{\operatorname{argmin}} \left(X_{(\cdot)} - A\begin{pmatrix} \mu \\ \sigma \end{pmatrix}\right)^t B^{-1} \left(X_{(\cdot)} - A\begin{pmatrix} \mu \\ \sigma \end{pmatrix}\right)$$

The solution is given by $\begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} = (A^t B^{-1} A)^{-1} A^t B^{-1} X_{(\cdot)}$,

where $A^t B^{-1} A = \begin{pmatrix} 1^t B^{-1} 1 & 1^t B^{-1} \alpha \\ \alpha^t B^{-1} 1 & \alpha^t B^{-1} \alpha \end{pmatrix}$.

- Fact: If the distribution of the Z_i is symmetric (e.g. $\mathcal{N}(0,1)$ distributed) then $1^t B^{-1} \alpha = 0$; so that the matrix $A^t B^{-1} A$ is diagonal.
- proof: follows that if the distribution of Z_i is symmetrical, then $(Z_{(1)}, \cdots, Z_{(n)}) \overset{d}{=} (-Z_{(n)}, \cdots, -Z_{(1)})$ (Balakrishnan & Cohen (1991)).

Let $J := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. It follows that

$\alpha = -J\alpha$ & $B = JBJ$

Mean of $Z_{(\cdot)}$ .. of $-JZ_{(\cdot)}$ — Cov of $Z_{(\cdot)}$ .. of $-JZ_{(\cdot)}$ (since $J^{-1} = J^t = J$)

$$\Rightarrow 1^t B^{-1} \alpha = 1^t (JB^{-1}J)(-J\alpha)$$

($\uparrow$ since $B^{-1} = JB^{-1}J$)

$$= -(1^t J) J^2 \alpha$$

($1^t J = 1^t$, $J^2 = I$)

$$= -1^t B^{-1} \alpha$$

$\Rightarrow$ Necessarily $1^t B^{-1} \alpha = 0$. ∎

× Summary = If the distribution of the Z_i is symmetric (which will be true in particular under the null hypothesis that the $X_1, \cdots, X_n$ are normally distributed), then

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} = \begin{pmatrix} 1^t B^{-1} 1 & 0 \\ 0 & \alpha^t B^{-1} \alpha \end{pmatrix}^{-1} \begin{pmatrix} 1^t \\ \alpha^t \end{pmatrix} B^{-1} X_{(\cdot)}$$

$$\Rightarrow \boxed{\hat{\mu} = \frac{1^t B^{-1} X_{(\cdot)}}{1^t B^{-1} 1}} \quad \& \quad \boxed{\hat{\sigma} = \frac{\alpha^t B^{-1} X_{(\cdot)}}{\alpha^t B^{-1} \alpha}}$$

← an L-statistic.

The Gauss-Markov theorem establishes that $(\hat{\mu}, \hat{\sigma})$ BLUE (Best Linear Unbiased Estimator) of (μ, σ)

Best: smallest covariance amongst the class of unbiased estimators of (μ, σ)

Linear: linear combination of $X_{(\cdot)}$

The Shapiro-Wilk statistic compares $\hat{\sigma}^2$ with the empirical variance estimator $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. Consider the ratio =

$$\frac{\hat{\sigma}^2}{(n-1)^{-1} \sum (X_i - \bar{X})^2} = \frac{(n-1)(\alpha^t B^{-1} X_{(\cdot)})^2}{\sum (X_i - \bar{X})^2 (\alpha^t B^{-1} \alpha)^2}$$

$$= (n-1) \frac{(\alpha^t B^{-1} B^{-1} \alpha)}{(\alpha^t B^{-1} \alpha)^2} \underbrace{\frac{(\alpha^t B^{-1} X_{(\cdot)})^2}{\sum (X_i - \bar{X})^2 (\alpha^t B^{-1} B^{-1} \alpha)}}_{W}$$

the SHAPIRO-WILK statistic.

The W-statistic can be rewritten:

$$W = \frac{\left(a^t X_{(\cdot)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \quad \text{with} \quad a_i^t = \frac{\alpha^t B^{-1}}{\sqrt{\alpha^t B^{-1} B^{-1} \alpha}}$$

SHAPIRO-WILK STATISTIC

W is proportional to the ratio of $\hat{\sigma}^2$ with s^2. The reason why it is take "proportional to" and not "exactly equal to" the ratio $\hat{\sigma}^2/s^2$ is that W benefits from nice properties, such as:

(i) $\sum_{i=1}^{n} a_i = 0$ for $a = (a_1, \ldots, a_n)^t$

(ii) W is scale & translation invariant

[Taking $X'_{(\cdot)} = \gamma X_{(\cdot)} + \delta$, the W statistic constructed from $X_{(\cdot)}$ and $X'_{(\cdot)}$ is the same]

- Consequence: suitable statistic for testing a composite hypothesis of normality.
- proof: follows easily from the definition of W, and the fact that $\sum_{i=1}^{n} a_i = 0$.

(iii) $W \leq 1$

- proof: note that $\sum_{i=1}^{n} a_i^2 = a^t a = 1$, so that the nominator of W is $\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2 \leq \left(\sum a_i^2\right)\left(\sum X_{(i)}^2\right) = \sum X_{(i)}^2$

Moreover, since the W statistic is translation invariant, we can assume without loss of generality that $\bar{X} = 0$ (otherwise consider $X_i \leftarrow X_i - \bar{X}$), so that the W-statistic can be reduced to

$$W = \frac{\left(\sum a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} X_i^2} \leq \frac{\sum X_{(i)}^2}{\sum X_i^2} = 1,$$

with equality if and only if $X_{(i)} \propto a_i$ (Cauchy-Schwartz inequality).

Summary: $W \leq 1$, with equality iff $\exists \lambda$ s.t. $X_{(i)} = \lambda a_i \quad \forall i = 1, \ldots, n$.

The coefficient W is a measure of linear association between the $X_{(i)}$ and the a_i. In fact, W is the square empirical (Pearson) correlation coefficient between $X_{(1)}, \ldots, X_{(n)}$ and $a_1, \ldots, a_n$:

$$W = \frac{\left(\sum a_i X_{(i)}\right)^2}{\sum (x_i - \bar{x})^2} = \frac{\left(\sum a_i X_{(i)} - \bar{X} \sum a_i\right)^2}{\sum (x_i - \bar{x})^2 \sum a_i^2}$$

since $\sum a_i = 0$ & $\sum a_i^2 = 1$

$$= \frac{\left(\sum a_i (X_{(i)} - \bar{X})\right)^2}{\sum (x_i - \bar{x})^2 \sum (a_i - \bar{a})^2}$$

$$= \frac{\left(\sum (X_{(i)} - \bar{X})(a_i - \bar{a})\right)^2}{\sum (x_i - \bar{x})^2 \sum (a_i - \bar{a})^2}$$

= square empirical correlation coefficient.

- Under the null, the $X_1, \dots, X_n$ are normally distributed, and we can write: $\boxed{X_{(i)} = \mu + \sigma \alpha_i + \varepsilon_i}$, (page 12) where the α_i correspond to the mean value of the ordered sample $Z_{(1)} \leq \dots \leq Z_{(n)}$, where $Z_i \sim \mathcal{N}(0,1)$ iid.

In other words, $\boxed{X_{(\cdot)} \overset{\approx}{\uparrow} \mu + \sigma \alpha}$ almost linear relation between $X_{(\cdot)}$ & α

↳ But a is proportional to $B^{-1}\alpha$ (page 15). Thus

$\Leftrightarrow$ $X_{(\cdot)}$ depends linearly on a

$\Leftrightarrow$ W-statistic is close to 1: a value of the Shapiro-Wilk statistic close to 1 is an indication that the data originates from a normal distribution. A value much smaller than 1 indicates a departure from the normal distribution.

How small W should be to reject the null, i.e. reject the hypothesis of normality depends on the distribution of W. Unfortunately, according to Shapiro & Wilk (1965), there is no explicit expression for $n \geq 4$. Critical values may be obtained using Monte-Carlo simulations.

Requires the knowledge of α and B for a normal sample. These values, and that of α, are tabulated, as a function of n.

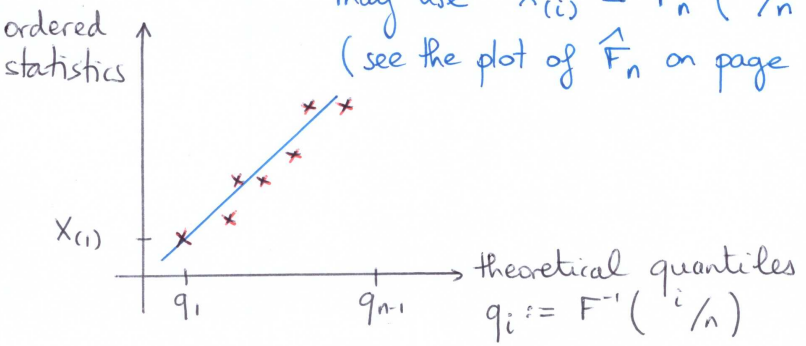
II.3. QQ plots.

QQ plots provide a visual way to perform goodness of fit tests.

Idea = check graphically if the plot of the EDF $\hat{F}_n$ is close to that of F; or equivalently, if the two inverses $\hat{F}_n^{-1}$ and F^{-1} are close to each other. Comparisons are made on a grid of values $1/n, \dots, (n-1)/n$

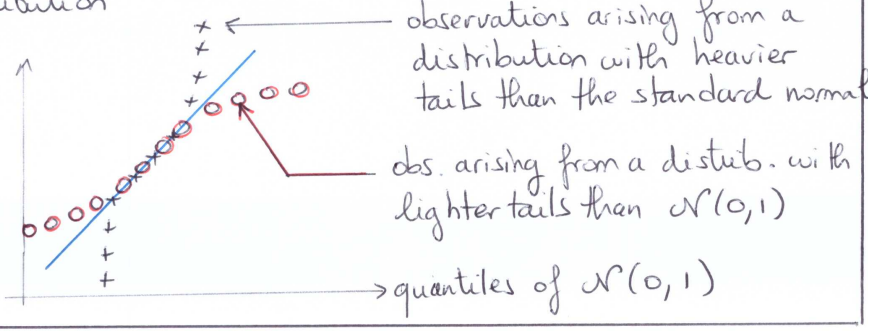
$\Rightarrow$ Check if the points $\left(F^{-1}\left(\frac{1}{n}\right), \hat{F}_n^{-1}\left(\frac{1}{n}\right)\right), \dots, \left(F^{-1}\left(\frac{n-1}{n}\right), \hat{F}_n^{-1}\left(\frac{n-1}{n}\right)\right)$ are close to the line $y = x$.

↖ the EDF is not invertible, but we may use $X_{(i)} = \hat{F}_n^{-1}(i/n)$ (see the plot of $\hat{F}_n$ on page 1).



Departure from $y = x$ indicates a departure from the distribution

Ex:

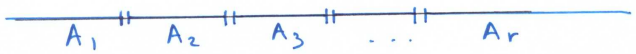


II.4. χ^2 goodness-of-fit test.

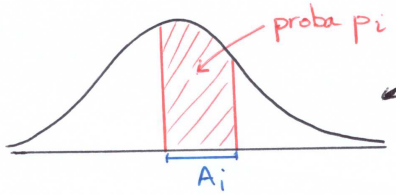
Let $X_1, \dots, X_n$ be iid RVs taking values in some set E (discrete or continuous).
Partition E into r bins $E = A_1 \cup \dots \cup A_r$.

Ex: (i) E is discrete; X take values in $E = \{a_1, \dots, a_r\}$.
Take $A_i = \{a_i\}$.

(ii) $E = \mathbb{R}$.



Suppose we only observe how many of the $X_1, \dots, X_n$ fall into each bin (and not X_i itself — in the discrete case, this is the same however). Let O_i denote the number of observations falling into bin A_i. Under the null hypothesis H_0, $X_1, \dots, X_n$ are assumed to arise from some distribution F, and we denote by E_i the expected number of observations falling into A_i, under H_0.



The probability that $X \sim F$ falls into A_i is $p_i = \int_{A_i} f(u)\,du$.
= Bernoulli trial with probability of success p_i.

$\Rightarrow$ Out of n observations, the number of observations falling into A_i is $B_i(n, p_i)$ distributed, so that the expected number of observations falling into A_i is $E_i = n p_i$.

Idea: Compare O_i and E_i. The closer they are, the less you want to reject H_0. To quantify this, we need to construct a statistic, whose distribution is known under H_0.

- Consider the simplest case: split E into two classes A_1 and A_2, so that $O_1 + O_2 = n$, where $O_i \sim B_i(n, p_i)$ (but not independent). We have $\mathbb{E} O_1 = n p_1$ & $\mathrm{Var}\, O_1 = n p_1 (1-p_1)$.

$$\text{CLT} \Longrightarrow \frac{O_1 - n p_1}{\sqrt{n p_1 (1-p_1)}} \approx \mathcal{N}(0,1) \Rightarrow \frac{(O_1 - n p_1)^2}{n p_1 (1-p_1)} \approx \chi^2(1)$$

Moreover,

$$\frac{(O_1 - n p_i)^2}{n p_1 (1-p_1)} = \frac{(O_1 - n p_1)^2 (1-p_1) + (O_1 - n p_1)^2 p_1}{n p_1 (1-p_1)}$$

$$= \frac{(O_1 - n p_1)^2}{n p_1} + \frac{(O_1 - n p_1)^2}{n(1-p_1)} \qquad O_1 = n - O_2$$

$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(-O_2 + n(1-p_1))^2}{n(1-p_1)}$$

$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \sim \chi^2(1) \quad (2 \text{ bins}), \quad = 2 - 1$$

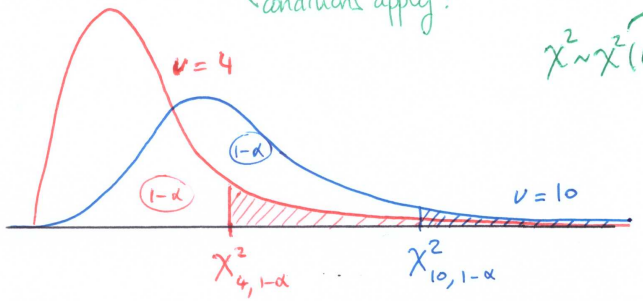
↖ Our statistic! Compare observed & expected number of observations falling into each bin, and sum all terms.

For r bins, take $\boxed{\chi^2 := \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i}}$. And we can show that under H_0, χ^2 has a $\chi^2(r-1)$ distribution, provided no parameters were estimated. If the null distribution F

contains d parameters estimated from the same dataset, the degrees of freedom must be adjusted to $r-d-1$, so that $\chi^2 \sim \chi^2(r-d-1)$ under H_0. Put $\alpha := \mathbb{P}(\chi^2 \leq \chi^2_{\nu,\alpha})$

↑ conditions apply!

$\chi^2 \sim \chi^2(\nu)$ ← Critical value



At a pre-specified confidence level $(1-\alpha)$, the critical value $\chi^2_{\nu,1-\alpha}$ decreases as the number of degrees of freedom decreased.

⇒ When estimating the parameters of the null distribution, we are using the data twice: for estimation purposes & to conduct the goodness-of-fit term. Therefore, the critical value is decreased, so that more evidence must be contained in the data in order not to reject the null.

× **Example**: Take $E = \mathbb{N}$, and $H_0 : \{P(\lambda)\}_{\lambda>0}$.

If one expects λ to be no larger than some λ_{max}, one can choose $A_1 = \{0\}$, $A_1 = \{1\}$, ..., $A_{r-1} = \{r-1\}$, and $A_r = \{r, r+1, \dots\}$, with r large enough such that $\mathbb{P}_{\lambda_{max}}(X \in A_r) \approx 0$. ∎

proof of $\chi^2 \sim \chi^2(r-1)$ [when no parameters are estimated]

- The random variables $\mathbb{1}(X_1 \in A_j), \dots, \mathbb{1}(X_n \in A_j)$ are iid $B(p_j)$, with mean p_j and variance $(1-p_j)$.

  CLT ⇒ $\dfrac{O_j - np_j}{\sqrt{np_j(1-p_j)}} \xrightarrow{d} \mathcal{N}(0,1)$,

  so that

  $\dfrac{O_j - np_j}{\sqrt{np_j}} \xrightarrow{d} \sqrt{1-p_j}\,\mathcal{N}(0,1) = \mathcal{N}(0, 1-p_j)$.

  Let's write $\boxed{\dfrac{O_j - np_j}{\sqrt{np_j}} \xrightarrow{d} Z_j \sim \mathcal{N}(0, 1-p_j).}$

  Unfortunately, the $Z_1, \dots, Z_r$ are correlated, so we cannot immediately deduce the distribution of the sum $Z_1^2 + \dots + Z_r^2$.

- We calculate the covariance between $\dfrac{O_i - np_i}{\sqrt{np_i}}$ and $\dfrac{O_j - np_j}{\sqrt{np_j}}$.

  $\mathrm{Cor}\left(\dfrac{O_i - np_i}{\sqrt{np_i}}, \dfrac{O_j - np_j}{\sqrt{np_j}}\right) = \dfrac{1}{n\sqrt{p_i p_j}}\,\mathbb{E}\{(O_i - np_i)(O_j - np_j)\}$

  ↑ zero mean

  $\mathbb{E}\{O_i O_j\} = n(n-1)\,p_i p_j$

  $= -\sqrt{p_i p_j}$

- **Summary**, $\displaystyle\sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{d} \sum_{i=1}^{r} Z_i^2$, $Z_i \sim \mathcal{N}(0, 1-p_i)$

  where $\begin{cases} \mathbb{E}(Z_i^2) = 1 - p_i \\ \mathbb{E}(Z_i Z_j) = -\sqrt{p_i p_j} \end{cases}$

- To obtain the distribution of $\sum_{i=1}^{r} z_i^2$, we find a different representation for it.

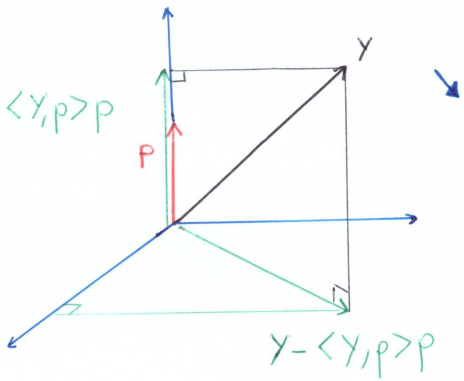
Consider $Y_1, \dots, Y_r$ iid $\mathcal{N}(0,1)$

$$\underset{(r\times 1)}{Y} := \begin{pmatrix} Y_1 \\ \vdots \\ Y_r \end{pmatrix} \qquad \underset{(r\times 1)}{p} := \begin{pmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_r} \end{pmatrix}$$

Claim: $Y - \langle Y, p\rangle p \overset{d}{=} \begin{pmatrix} z_1 \\ \vdots \\ z_r \end{pmatrix}$

To see this, check that their mean & covariance matrices (i.e. second order properties) are the same.

- **Geometrical considerations:**
  - $\langle Y, p\rangle p$ = orthogonal projection of Y onto the subspace of dimension 1, in the direction of p
  - $Y - \langle Y, p\rangle p$ = projection onto the orthogonal complement, of dimension $r-1$.
    - ↳ COCHRAN theorem (see p.24 in SL = LINEAR REGRESSION) ensures that $\| Y - \langle Y, p\rangle p \|^2 \sim \chi^2(r-1)$
    - $r-1$ = dimension of the subspace



⇒ We conclude that indeed $\sum_{i=1}^{r} z_i^2 \sim \chi^2(r-1)$. ∎

III - TESTS ON 2 SAMPLES

III.1. Kolmogorov - Smirnov test

Suppose that $\begin{cases} X_1, \dots, X_{n_1} \sim F \\ Y_1, \dots, Y_{n_2} \sim G \end{cases}$

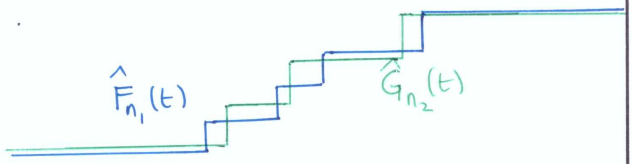
We want to test $H_0 : F = G$.

The Kolmogorov - Smirnov test, based on EDFs, is very similar to the one when testing 1 sample. (p.7)

Denote $\begin{cases} \hat{F}_{n_1}(t) = \text{EDF constructed from } X_1, \dots, X_{n_1} \\ \hat{G}_{n_2}(t) = \text{EDF} \quad \text{——"——} \quad Y_1, \dots, Y_{n_2} \end{cases}$

Consider the maximum difference D_{n_1,n_2} observed between the two EDFs: $\boxed{D_{n_1,n_2} := \sup_t | \hat{F}_{n_1}(t) - \hat{G}_{n_2}(t) |}$.

The challenge is to derive the distribution of D_{n_1,n_2} under the null H_0.



The good news is that here as well, the distribution of D_{n_1,n_2} does not depend on $F = G$ (under H_0), since

$$D_{n_1,n_2} = \sup_t | \hat{u}_{n_1}(F(t)) - \hat{u}_{n_2}(F(t)) | = \sup_{u\in[0,1]} | \hat{u}_{n_1}(u) - \hat{u}_{n_2}(u) |.$$

(same notation as on page 3.)

Moreover, it can be shown that

$$\lim_{\substack{n_1\to\infty \\ n_2\to\infty}} \mathbb{P}\left(\sqrt{\frac{n_1 n_2}{n_1+n_2}}\, D_{n_1,n_2} \le x \right) = K(x), \quad x>0.$$

(K: Kolmogorov function (p.7))

III.2. Wilcoxon-Mann-Whitney test

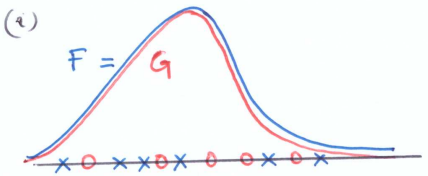
Let $X_1, \dots, X_{n_1} \sim F$ iid size n_1
$Y_1, \dots, Y_{n_2} \sim G$ iid size n_2

Test for $H_0: F = G$

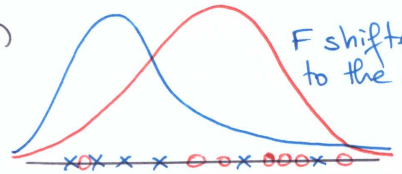
Assume that F and G are absolutely continuous, so that we have no ties (with probability 1).

Idea: Order the combined sample $\{X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}\}$ of size n_2. Under H_0, observations from the first and second sample should alternate regularly. Compare:

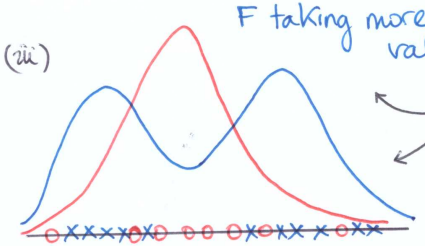
(i) $F = G$



(ii) F shifted to the right



(iii) F taking more extreme values



In (i), $F = G$ and observations x and o alternate. In (ii) and (iii), observations x (resp. o) tend to group together.

⇒ The Mann-Whitney statistic computes the number of times observations from one sample precedes observations of the other sample.

× Drawback: powerful when one distribution is shifted with respect to the other; but not if the centers of mass are close to each other.

↳ For each X_i, compute the number of observations from $\{Y_1, \dots, Y_{n_2}\}$ smaller than it. It is given by $\sum_{j=1}^{n_2} \mathbb{1}(Y_j < X_i)$.

Denote by U_X the total number of times we have that $X_i > Y_j$:

$$U_X = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbb{1}(Y_j < X_i)$$

Likewise, define

$$U_Y = \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \mathbb{1}(X_i < Y_j)$$

MANN-WHITNEY.

↳ Note that

$$U_X + U_Y = \sum_{i,j} \{ \underbrace{\mathbb{1}(Y_j < X_i) + \mathbb{1}(X_i < Y_j)}_{=1 \text{ (assuming no ties)}} \} = n_1 n_2.$$

⇒ Only need to compute U_X or U_Y.

↳ We have

$$R_{X_i} = \sum_{j=1}^{n_2} \mathbb{1}(Y_j < X_i) + \sum_{k=1}^{n_1} \mathbb{1}(X_k < X_i)$$

↑ Rank of X_i in the combine sample = Sum all observations (coming from sample 1 & 2) less than X_i.

$$\Rightarrow U_X = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbb{1}(Y_j < X_i)$$

$$= \sum_{i=1}^{n_1} \{ R_{X_i} - \sum_{k=1}^{n_1} \mathbb{1}(X_k < X_i) \}$$

$$= \sum_{i=1}^{n_1} R_{X_i} - \sum_{i=1}^{n_1} \sum_{k=1}^{n_1} \mathbb{1}(X_k < X_i)$$

$$\| \quad R_X = \text{sum of the ranks of the first sample.}$$

In addition, $\sum_{i=1}^{n_1} \sum_{k=1}^{n_1} \mathbb{1}(X_k < X_i)$ corresponds to the sum of the ranks for the first sample only. It is equal to $1 + 2 + \cdots + n_1 = \frac{n_1(n_1+1)}{2}$.

Thus, $\boxed{U_X = R_X - \frac{n_1(n_1+1)}{2}}$.

Likewise, $\boxed{U_Y = R_Y - \frac{n_2(n_2+1)}{2}}$.

The Mann-Whitney U-statistic

The Wilcoxon Rank-Sum statistic

Under H_0, the distribution of U_X / R_X (resp. U_Y / R_Y) is independent of $F = G$. Indeed, it follows from symmetry considerations that

$$\mathbb{P}\left(R_{X_1} = i_1, \cdots, R_{X_{n_1}} = i_{n_1}\right) = \frac{1}{\binom{n_1+n_2}{n_1}},$$

for $1 \leq i_1 < \cdots < i_{n_1} \leq n_1 + n_2$.

For small values of n_1/n_2, this distribution is tabulated. For n_1/n_2 large enough, the normal approximation may be used. We have

$$\mathbb{E}\, U_X = \sum_{i,j} \underbrace{\mathbb{P}(Y_j < X_i)}_{= 1/2 \text{ under } H_0} = \frac{n_1 n_2}{2}$$

$$\mathrm{Var}\, U_X = \mathrm{Var}\left\{ \sum_{i,j} \mathbb{1}(Y_j < X_i) \right\} = \cdots = \frac{n_1 n_2 (n_1+n_2+1)}{12}$$

↑ requires more work as the variables are correlated.

$$\Rightarrow \frac{U_X - n_1 n_2/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}} \approx \mathcal{N}(0,1)$$

↑ Use the normal approximation to define a rejection region. — Business as usual —

(Alternatively, use $\mathbb{E} R_X = \mathbb{E} U_X + \frac{n_1(n_1+1)}{2} = \frac{n_1(n_1+n_2+1)}{2}$, and the normal approximation $\frac{R_X - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}} \approx \mathcal{N}(0,1)$).

[Ref] E.L. Lehmann. Elements of Large-Sample-Theory. Springer. (p.147/148).