Statistical Frontiers in Functional Data: A Spotlight on ANOVA Techniques

Nuri Celik

Gebze Technical University

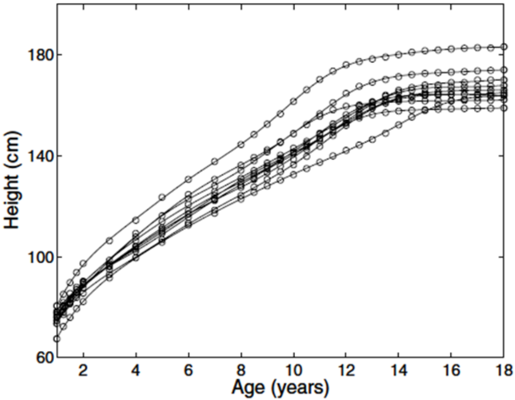
ncelik@gtu.edu.tr

November 26, 2024

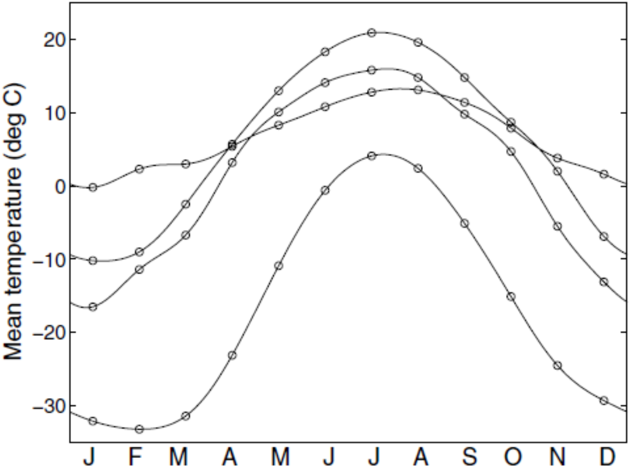
Presentation Overview

Functional Data

Source: Jin-Ting Zhang, Analysis of Variance for Functional Data, 2014



The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.



Mean monthly temperatures for the Canadian weather stations. In descending order of the temperatures at the start of the year, the stations are Prince Rupert, Montreal, Edmonton, and Resolute.

- To represent the data in ways that aid further analysis
- To display the data so as to highlight various characteristics
- To study important sources of pattern and variation among the data
- To explain variation in an outcome or dependent variable by using input or independent variable information

The advantages of FDA over Time Series, Multivariate Data Analysis and Repeated Measures

- Continuity of data
- Less information loss
- Capturing complex structures
- Solution for high dimensionality
- Flexibility in Handling Unevenly Spaced or Missing Data
- Practical Application in Real-World Contexts

Spline Methodology

A functional data set

$$(t_i, y_i), \ i = 1, 2, \ldots, n,$$

- Data Fitting: To fit the data with a smooth curve.
- Function Approximation: For approximating complex functions.
- Curve Shaping: To ensure specific behaviors at certain points on the curve.

Some common types of splines include:

- Linear Splines
- Cubic Splines
- B-Splines

Given data points:

$$(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$$

For instance:

$$(0, 1), (1, 2), (2, 0), (3, 2), (4, 3)$$

Let's define a cubic polynomial for each interval:

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \quad i = 0, 1, \ldots, n - 1$$

Here, $S_i(x)$ is the polynomial in the i-th interval (i.e., $x_i \leq x \leq x_{i+1}$).

To determine the splines, the following conditions must be satisfied: 1. Fitting Condition: The spline curve should pass through each data point:

$$S_i(x_i) = y_i \quad \text{and} \quad S_i(x_{i+1}) = y_{i+1}$$

2. First Derivative Continuity: The first derivative of the curve should be continuous across intervals:

$$S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$$

3. Second Derivative Continuity: The second derivative of the curve should be continuous across intervals:

$$S_i''(x_{i+1}) = S_{i+1}''(x_{i+1})$$

4. Natural Spline Boundary Conditions (Optional): The second derivatives at the endpoints should be zero:

$$S_0''(x_0) = 0 \quad \text{and} \quad S_{n-1}''(x_n) = 0$$

Using these conditions, the coefficients a_i, b_i, c_i, and d_i are determined.

In a smoothed spline, the curve does not pass exactly through all the data points. Instead, a balance parameter (λ) is used to minimize the following function:

$$\sum_{i=0}^{n}(y_i - S(x_i))^2 + \lambda \int_{x_0}^{x_n} (S''(x))^2 dx$$

Where: - $(y_i - S(x_i))^2$ measures the fit of the spline to the data. - $\int_{x_0}^{x_n}(S''(x))^2 dx$ measures the smoothness of the curve. - λ controls the balance between fitting the data and the smoothness of the curve. A small λ value results in a curve that fits the data closely, while a large λ value yields a smoother curve.

The degree of smoothing of the y_j measurement values is decided by considering the number of basis functions (K). For the desired x function, a basis system consisting of K basis functions is selected. These basis functions are denoted by $\phi_k, k = 1, \ldots, K$. The functional data to be generated is obtained from the linear combination of the basis functions:

$$y(t) = \sum_{k=1}^{K} c_k \phi_k(t) = c' \phi(t).$$

The $c_1, c_2, \ldots, c_k$ parameters are the coefficients that determine the shape and form of the function corresponding to this basis function. Considering the functional data to be created for N individuals, it can be represented as:

$$y_i(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t), \quad i = 1, 2, \ldots, N.$$

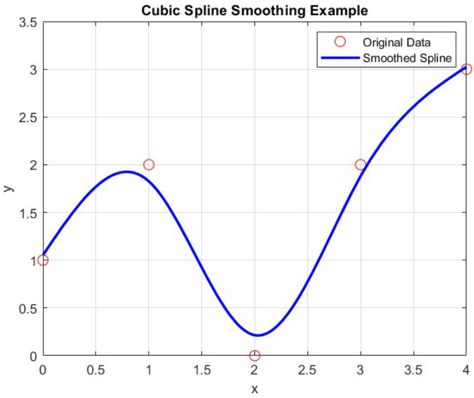


Figure: Example of Functional Data.

Smoothing parameter (lambda): 0.0100
Piecewise polynomial coefficients: Interval [0.000000, 1.000000]:
-0.881708 (x^3) $+ 0.000000(x^2) + 1.654376(x) +$
$1.052902 Interval[1.000000, 2.000000]:$
$2.025448(x^3) + -2.645124(x^2) + -0.990748(x) +$
$1.825571 Interval[2.000000, 3.000000]:$
$-1.560332(x^3) + 3.431221(x^2) + -0.204651(x) +$
$0.215147 Interval[3.000000, 4.000000]:$
$0.416592(x^3) + -1.249775(x^2) + 1.976795(x) + 1.881385$

Descriptive Statistics

Functional Mean

$$x(t) = N^{-1} \sum_{i=1}^{N} x_i(t)$$

Functional Variance

$$\text{var} X(t) = (N-1)^{-1} \sum_{i=1}^{N} [x_i(t) - \bar{x}(t)]^2,$$

Covariance Function

$$\text{cov} X(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} \{x_i(t_1) - \bar{x}(t_1)\}\{x_i(t_2) - \bar{x}(t_2)\}.$$

Stochastic Process (SP)

A stochastic process is a collection of random variables indexed by time or space, often written as:

$$\{y(t) : t \in T\},$$

We write $y(t) \sim SP(\eta, \gamma)$ for simplicity. When $\gamma(s, t)$ has finite trace:

$$\text{tr}(\gamma) = \int_T \gamma(t, t)dt < \infty,$$

it has the following singular value decomposition (SVD) :

$$\gamma(s, t) = \sum_{r=1}^{m} \lambda_r \phi_r(s)\phi_r(t),$$

where $\lambda_1, \lambda_2, \ldots, \lambda_m$ are all the decreasingly ordered positive eigenvalues of $\gamma(s, t)$, and $\phi_1(t), \phi_2(t), \ldots, \phi_m(t)$ are the associated orthonormal eigenfunctions.

As the covariance function $\gamma(s,t)$ is symmetric with respect to s and t, we have

$$\mathrm{tr}(\gamma^{\otimes 2}) = \iint_T \gamma^2(s,t)\, ds\, dt.$$

By the SVD of $\gamma(s,t)$, we can show that

$$\mathrm{tr}(\gamma) = \sum_{r=1}^{m} \lambda_r, \quad \mathrm{tr}(\gamma^{\otimes 2}) = \sum_{r=1}^{m} \lambda_r^2.$$

Gaussian Process

A process $y(t), t \in T$ is Gaussian with mean function $\eta(t), t \in T$ and covariance function $\gamma(s, t), s, t \in T$, denoted as $GP(\eta, \gamma)$, if and only if for any p time points, $t_j, j = 1, 2, \ldots, p$, the random vector

$$[y(t_1), \ldots, y(t_p)]^T$$

follows a multivariate normal distribution $N_p(\eta, \Gamma)$, where

$$\eta = [\eta(t_1), \ldots, \eta(t_p)]^T$$

and

$$\Gamma = (\gamma(t_i, t_j)) : p \times p.$$

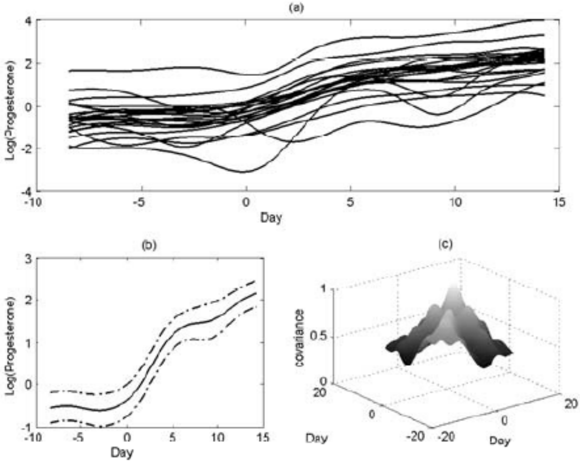
Wishart Process

Wishart processes are natural generalizations of Wishart random matrices. Throughout, we use $WP(n, \gamma)$ to denote a Wishart process with n degrees of freedom and a covariance function $\gamma(s, t)$. A general Wishart process $W(s, t) \sim WP(n, \gamma)$ can be written as

$$W(s, t) = \sum_{i=1}^{n} W_i(s, t) = \sum_{i=1}^{n} v_i(s) v_i(t),$$

where $W_i(s, t) = v_i(s) v_i(t)$, $i = 1, 2, \ldots, n$, $i \overset{\text{i.i.d.}}{\sim} WP(1, \gamma)$ and $v_i(t)$, $i = 1, 2, \ldots, n$, $i \overset{\text{i.i.d.}}{\sim} GP(0, \gamma)$.

One-sample Problem



The conceptive progesterone data example with (a) reconstructive individual curves, (b) sample mean function with its 95% pointwise confidence bands, and (c) sample covariance function.

$$H_0 : \eta(t) \equiv -0.50,\ t \in [a, b],$$

versus

$$H_1 : \eta(t) \neq -0.50,\ \text{for some } t \in [a, b],$$

where $[a, b]$ is any time period of interest. When $[a, b] = [-8, 0]$, $[0, 15]$, and $[-8, 15]$.

$$y_1(t), \ldots, y_n(t) \overset{\text{i.i.d.}}{\sim} SP(\eta, \gamma),$$

and we wish to test the following hypothesis testing problem:

$$H_0 : \eta(t) \equiv \eta_0(t),\ t \in T, \quad \text{versus} \quad H_1 : \eta(t) \neq \eta_0(t),\ \text{for some } t \in T.$$

One-Sample Problem Assumptions (OS)

1. The functional sample is with $\eta(t) \in L^2(T)$ and $\text{tr}(\gamma) < \infty$.
2. The functional sample is Gaussian.
3. The subject-effect function $v_1(t)$ satisfies $\mathbb{E}||v_1||^4 = \mathbb{E}\left[\left(\int_T v_1^2(t)\,dt\right)\right]^2 < \infty$.
4. The maximum variance $\rho = \max_{t \in T} \gamma(t,t) < \infty$.

Theorem

Under Assumptions OS1 and OS2, we have

$$\sqrt{n}\{\hat{\eta}(t) - \eta(t)\} \sim GP(0, \gamma), \quad (n-1)\hat{\gamma}(s,t) \sim WP(n-1, \gamma).$$

Theorem

Under Assumption OS1, as $n \to \infty$, we have

$$\sqrt{n}\{\hat{\eta}(t) - \eta(t)\} \xrightarrow{d} GP(0, \gamma),$$

Theorem

Under Assumptions OS1, OS3, and OS4, as $n \to \infty$, we have

$$\sqrt{n}\{\hat{\gamma}(s,t) - \gamma(s,t)\} \xrightarrow{d} GP(0, \Phi),$$

where

$$\Phi\{(s_1,t_1),(s_2,t_2)\} = \mathbb{E}\{v_1(s_1)v_1(t_1)v_1(s_2)v_1(t_2)\} - \gamma(s_1,t_1)\gamma(s_2,t_2).$$

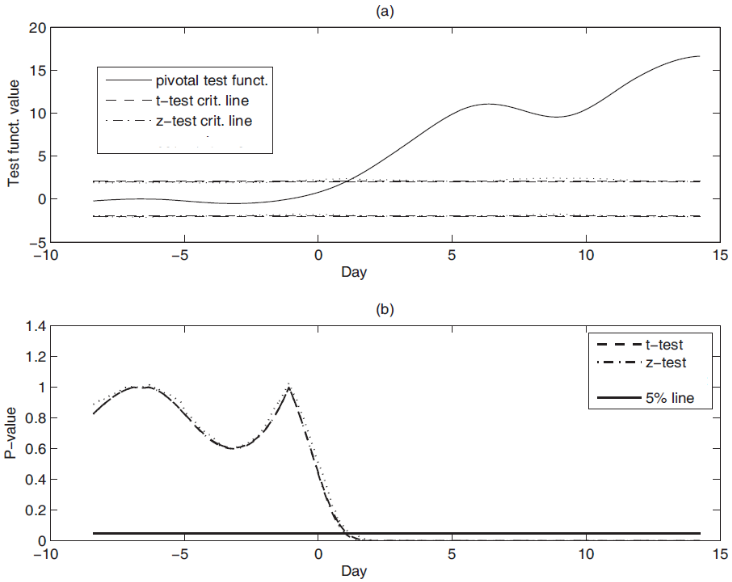
Pointwise Test

Pointwise tests aim to test the null hypothesis in the one-sample problem at each time point $t \in T$. For any fixed t, the sub-problem is

$$H_{0t} : \eta(t) = \eta_0(t), \quad \text{versus} \quad H_{1t} : \eta(t) \neq \eta_0(t).$$

Based on the estimators, the pivotal test statistic for the above local hypothesis testing problem is

$$z(t) = \frac{\Delta(t)}{\sqrt{\hat{\gamma}(t,t)}} = \frac{\sqrt{n}[\bar{y}(t) - \eta_0(t)]}{\sqrt{\hat{\gamma}(t,t)}}.$$


(a)
pivotal test funct.
t–test crit. line
z–test crit. line
Test funct. value
Day
(b)
t–test
z–test
5% line
P–value
Day

L^2-Norm Based Test

Under H_0, when the Gaussian assumption is valid, the pivotal test function $\Delta(t) \sim \text{GP}(0, \gamma)$ and when the Gaussian assumption is not valid but n is large, we have $\Delta(t) \xrightarrow{d} \text{GP}(0, \gamma)$. The L^2-norm-based test uses the squared L^2-norm of $\Delta(t)$ as its test statistic:

$$T_n = \|\Delta\|^2 = n \int_T [\bar{y}(t) - \eta_0(t)]^2 \, dt.$$

It is easy to see that T_n will be small under the null hypothesis and it will be large under the alternatives. and under H_0,

$$T_n \stackrel{d}{=} \sum_{r=1}^{m} \lambda_r A_r, \quad A_r \stackrel{\text{i.i.d.}}{\sim} \chi_1^2,$$

which is valid when the Gaussian assumption holds or is asymptotically valid when n is large.

The null distribution of T_n can then be approximated by the Welch-Satterthwaite χ^2-approximation, we have

$$T_n \sim \beta\chi^2_d \text{ approximately}, \quad \text{where} \quad \beta = \frac{\text{tr}(\gamma^{\otimes 2})}{\text{tr}(\gamma)}, \quad d = \kappa = \frac{\text{tr}^2(\gamma)}{\text{tr}(\gamma^{\otimes 2})}.$$

F Type Test

For the one-sample problem, recall that the L^2-norm-based test is based on $\|\Delta\|^2$ with the pivotal test function $\Delta(t) = \sqrt{n}[\bar{y}(t) - \eta_0(t)]$. Under H_0 and the Gaussian assumption, $\Delta(t) \sim \text{GP}(0, \gamma)$ and $(n-1)\hat{\gamma}(s,t) \sim \text{WP}(n-1, \gamma)$ are independent. In addition,

$$\mathbb{E}[\|\Delta\|^2] = \text{tr}(\gamma), \quad \mathbb{E}[\text{tr}(\hat{\gamma})] = \text{tr}(\gamma).$$

Then it is natural to test using the following test statistic:

$$F_n = \frac{\|\Delta\|^2}{\text{tr}(\hat{\gamma})} = \frac{n\int_T [\bar{y}(t) - \eta_0(t)]^2 dt}{\text{tr}(\hat{\gamma})}.$$

where

$$\|\Delta\|^2 \overset{d}{=} \sum_{r=1}^{m} \lambda_r A_r, \quad A_r \overset{\text{i.i.d.}}{\sim} \chi_1^2,$$

$$\text{tr}(\hat{\gamma}) \overset{d}{=} \sum_{r=1}^{m} \lambda_r E_r, \quad E_r \overset{\text{i.i.d.}}{\sim} \chi_{n-1}^2,$$

Two Sample Problem

A general two-sample problem for functional data with a common covariance function can be formulated as follows. Suppose we have two functional samples

$$y_{11}(t), \cdots, y_{1n_1}(t) \overset{i.i.d.}{\sim} SP(\eta_1, \gamma), \quad y_{21}(t), \cdots, y_{2n_2}(t) \overset{i.i.d.}{\sim} SP(\eta_2, \gamma),$$

where $\eta_1(t)$ and $\eta_2(t)$ are the unknown mean functions of the two samples, and $\gamma(s, t)$ is their common covariance function, which is usually unknown. We wish to test the following hypotheses:

$$H_0 : \eta_1(t) \equiv \eta_2(t), \quad t \in T,$$

versus

$$H_1 : \eta_1(t) \neq \eta_2(t), \quad \text{for some } t \in T,$$

where T is the time period of interest, often a finite interval $[a, b]$ say with $-\infty < a < b < \infty$.

To test the two-sample problem based on the two samples , a natural pivotal test function is

$$\Delta(t) = \frac{n_1 n_2}{n} \left(\bar{y}_1(t) - \bar{y}_2(t)\right),$$

which is the scaled mean function difference of the two samples. When the null hypothesis is valid, this quantity will be small and it will be large otherwise. Therefore, it is appropriate to use $\Delta(t)$ as a pivotal test function for the two-sample problem . Notice that $\Delta(t)$ has its mean and covariance functions as

$$\eta_\Delta(t) = \mathbb{E}\Delta(t) = \frac{n_1 n_2}{n} \left(\eta_1(t) - \eta_2(t)\right),$$

and

$$\text{Cov}[\Delta(s), \Delta(t)] = \gamma(s, t).$$

Under the null hypothesis, we have $\mathbb{E}\Delta(t) \equiv 0$, $t \in T$.

Two-Sample Problem Assumptions (TS)

1. The two samples are with $\eta_1(t), \eta_2(t) \in L^2(T)$ and $\text{tr}(\gamma) < \infty$.
2. The two samples are Gaussian.
3. As $n \to \infty$, the sample sizes satisfy $n_1/n \to \tau$ such that $\tau \in (0, 1)$.
4. The subject-effect functions $V_{ij}(t) = y_{ij}(t) - \eta_i(t)$, $j = 1, 2, \cdots, n_i$; $i = 1, 2$ are i.i.d..
5. The subject-effect function $v_{11}(t)$ satisfies

$$\mathbb{E}||v_{11}||^4 = \mathbb{E}\left(\int_T v_{11}^2(t)\, dt\right)^2 < \infty.$$

6. The maximum variance $\rho = \max_{t \in T} \gamma(t, t) < \infty$.

Theorem

Under Assumptions TS1 and TS2, we have

$$\Delta(t) \sim GP(\eta_\Delta, \gamma),$$

and

$$(n-2)\hat{\gamma}(s,t) \sim WP(n-2, \gamma).$$

$$\sqrt{\frac{n_1 n_2}{n}}(\bar{y}_1(t) - \bar{y}_2(t)) \sim GP(0, \gamma).$$

Theorem

Under Assumptions TS1, TS3, and TS4, as $n \to \infty$, we have

$$\Delta(t) - \eta_\Delta(t) \xrightarrow{d} GP(0, \gamma),$$

Theorem

Under Assumptions TS1 and TS3 through TS6, as $n \to \infty$, we have

$$\sqrt{n}\{\hat{\gamma}(s,t) - \gamma(s,t)\} \xrightarrow{d} GP(0, \Phi),$$

where

$$\Phi\{(s_1, t_1), (s_2, t_2)\} = \mathbb{E}\{v_{11}(s_1)v_{11}(t_1)v_{11}(s_2)v_{11}(t_2)\} - \gamma(s_1, t_1)\gamma(s_2, t_2).$$

Pointwise Tests

We here describe pointwise t-, z-, and bootstrap tests for the two-sample problem under various conditions. The key idea of a pointwise test is to test the null hypothesis at each time point $t \in T$. For any fixed $t \in T$, the sub-problem is

$$H_{0t} : \eta_1(t) = \eta_2(t), \quad \text{versus} \quad H_{1t} : \eta_1(t) \neq \eta_2(t).$$

Based on the sample mean functions and the pooled sample covariance function given in (5.3), the pivotal test statistic is

$$z(t) = \frac{[\bar{y}_1(t) - \bar{y}_2(t)]}{\sqrt{(1/n_1 + 1/n_2)\hat{\gamma}(t,t)}} = \frac{\Delta(t)}{\sqrt{\hat{\gamma}(t,t)}},$$

where $\Delta(t)$ is the pivotal test function.

L^2-Norm Based Test

For the two-sample problem, the L^2-norm-based test uses the squared L^2-norm of the pivotal test function $\Delta(t)$ (5.4) as the test statistic:

$$T_n = \int_T \Delta^2(t)\, dt = \frac{n_1 n_2}{n} \int_T [\bar{y}_1(t) - \bar{y}_2(t)]^2 \, dt.$$

Under the null hypothesis, when the two functional samples are Gaussian, we have $\Delta(t) \sim GP(0, \gamma)$, and when the two samples are large and satisfy Assumptions TS1, TS3, and TS4, we have $\Delta(t) \sim GP(0, \gamma)$ asymptotically. Therefore, we have

$$T_n \stackrel{d}{=} \sum_{r=1}^{m} \lambda_r A_r, \quad A_r \stackrel{i.i.d.}{\sim} \chi_1^2,$$

where $\lambda_1, \lambda_2, \cdots, \lambda_m$ are all the positive eigenvalues of the common covariance function $\gamma(s, t)$.

$$T_n \sim \beta \chi_d^2 \quad \text{approximately},$$

F Type Test

For the two-sample problem, recall that the L^2-norm-based test is based on the squared L^2-norm $\|\Delta\|^2$ of the pivotal test function

$$\Delta(t) = \frac{n_1 n_2}{n} \left(\bar{y}_1(t) - \bar{y}_2(t) \right).$$

Under the null hypothesis and the Gaussian assumption,

$$\Delta(t) \sim GP(0, \gamma) \quad \text{and} \quad (n-2)\hat{\gamma}(s,t) \sim WP(n-2, \gamma)$$

and they are independent. In addition, we have

$$\mathbb{E}[\|\Delta\|^2] = \text{tr}(\gamma), \quad \text{and} \quad \mathbb{E}[\text{tr}(\hat{\gamma})] = \text{tr}(\gamma).$$

Therefore, it is natural to test using the following F-type test statistic:

$$F_n = \frac{\|\Delta\|^2}{\text{tr}(\hat{\gamma})} = \frac{n_1 n_2}{n} \frac{\int_T [\bar{y}_1(t) - \bar{y}_2(t)]^2 \, dt}{\text{tr}(\hat{\gamma})}.$$

When the variation of $\mathrm{tr}(\hat{\gamma})$ is not taken into account, the distribution of F_n is essentially the same as that of the L^2-norm-based test statistic. To take this variation into account, the Gaussian assumption is sufficient. In fact, under the Gaussian assumption and the null hypothesis, by Theorem 4.2, we have

$$\|\Delta\|^2 \stackrel{d}{=} \sum_{r=1}^{m} \lambda_r A_r, \quad A_r \stackrel{i.i.d.}{\sim} \chi_1^2,$$

and

$$\mathrm{tr}(\hat{\gamma}) \stackrel{d}{=} \frac{\sum_{r=1}^{m} \lambda_r E_r}{n-2}, \quad E_r \stackrel{i.i.d.}{\sim} \chi_{n-2}^2,$$

where A_r, E_r are all independent and $\lambda_1, \lambda_2, \ldots, \lambda_m$ are all the positive eigenvalues of $\gamma(s,t)$. Equivalently, we can write

$$F_n \stackrel{d}{=} \frac{\sum_{r=1}^{m} \lambda_r A_r}{\sum_{r=1}^{m} \lambda_r E_r/(n-2)}.$$

One-way ANOVA

We can define the one-way ANOVA problem for functional data as follows. Suppose we have k independent samples:

$$y_{i1}(t), \ldots, y_{in_i}(t), \quad i = 1, \ldots, k.$$

These k samples satisfy

$$y_{ij}(t) = \eta_i(t) + v_{ij}(t), \quad v_{ij}(t) \overset{i.i.d.}{\sim} SP(0, \gamma),$$

where $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$, and $\eta_1(t), \eta_2(t), \ldots, \eta_k(t)$ are the unknown group mean functions of the k samples,
$v_{ij}(t)$, $j = 1, \ldots, n_i$; $i = 1, 2, \ldots, k$ are the subject-effect functions, and $\gamma(s, t)$ is the common covariance function. We wish to test the following one-way ANOVA testing problem:

$$H_0 : \eta_1(t) \equiv \eta_2(t) \equiv \ldots \equiv \eta_k(t), \quad t \in T.$$

Based on the k samples, the group mean functions $\eta_i(t),\ i = 1, 2, \ldots, k$ and the common covariance function $\gamma(s, t)$ can be unbiasedly estimated as

$$\hat{\eta}_i(t) = \bar{y}_{i.}(t) = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}(t), \quad i = 1, 2, \ldots, k,$$

$$\hat{\gamma}(s, t) = (n-k)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} [y_{ij}(s) - \bar{y}_{i.}(s)][y_{ij}(t) - \bar{y}_{i.}(t)],$$

where and throughout this section $n = \sum_{i=1}^{k} n_i$ denotes the total sample size. The estimated covariance function $\hat{\gamma}(s, t)$ is also known as the pooled sample covariance function. Note that $\hat{\eta}_i(t),\ i = 1, 2, \ldots, k$ are independent and

$$\mathbb{E}[\hat{\eta}_i(t)] = \eta_i(t), \quad \text{cov}\left(\hat{\eta}_i(s), \hat{\eta}_i(t)\right) = \frac{\gamma(s, t)}{n_i},\ i = 1, 2, \ldots, k.$$

Set $\hat{\eta}(t) = [\hat{\eta}_1(t), \hat{\eta}_2(t), \ldots, \hat{\eta}_k(t)]^T$. It is an unbiased estimator of $\eta(t)$. Then we have

$$\mathbb{E}[\hat{\eta}(t)] = \eta(t) \quad \text{and} \quad \text{Cov}\left(\hat{\eta}(s), \hat{\eta}(t)\right) = \gamma(s,t)D,$$

where $D = diag\left(\frac{1}{n_1}, \frac{1}{n_2}, \ldots, \frac{1}{n_k}\right)$ is a diagonal matrix with diagonal entries $\frac{1}{n_i}$, $i = 1, 2, \ldots, k$. That is, $\hat{\eta}(t) \sim SP_k(\eta, \gamma D)$, where $SP_k(\eta, \Gamma)$ denotes a k-dimensional stochastic process having the vector of mean functions $\eta(t)$ and the matrix of covariance functions $\Gamma(s,t)$.

One-Way ANOVA Assumptions (KS)

1. The k samples are with $\eta_1(t), \eta_2(t), \ldots, \eta_k(t) \in L^2(T)$ and $\text{tr}(\gamma) < \infty$.
2. The k samples are Gaussian.
3. As $n \to \infty$, the k sample sizes satisfy $\frac{n_i}{n} \to \tau_i$, $i = 1, 2, \ldots, k$ such that $\tau_1, \tau_2, \ldots, \tau_k \in (0, 1)$.
4. The subject-effect functions $v_{ij}(t) = y_{ij}(t) - \eta_i(t)$, $j = 1, 2, \ldots, n_i;\ i = 1, 2, \ldots, k$ are i.i.d.
5. The subject-effect function $v_{11}(t)$ satisfies $\mathbb{E}[|v_{11}|^4] < \infty$.
6. The maximum variance $\rho = \max_{t \in T} \gamma(t, t) < \infty$.
7. The expectation $\mathbb{E}[v_{11}^2(s) v_{11}^2(t)]$ is uniformly bounded.

Theorem

Under Assumptions KS1 and KS2, we have

$$D^{-1/2}[\hat{\eta}(t) - \eta(t)] \sim GP_k(0, \gamma I_k), \quad \text{and} \quad (n-k)\hat{\gamma}(s,t) \sim WP(n-k, \gamma).$$

Theorem

Under Assumptions KS1, KS3, and KS4, as $n \to \infty$, we have

$$D^{-1/2}[\hat{\eta}(t) - \eta(t)] \xrightarrow{d} GP_k(0, \gamma I_k).$$

Theorem

Under Assumptions KS1, KS3, KS4, KS5, and KS6, as $n \to \infty$, we have

$$\sqrt{n}\{\hat{\gamma}(s,t) - \gamma(s,t)\} \xrightarrow{d} GP(0, \mathcal{V}),$$

where

$$\mathcal{V}\{(s_1,t_1),(s_2,t_2)\} = \mathbb{E}\{v_{11}(s_1)v_{11}(t_1)v_{11}(s_2)v_{11}(t_2)\} - \gamma(s_1,t_1)\gamma(s_2,t_2).$$

Pointwise Test

For the main-effect, post hoc, or contrast tests, we do not need to identify the main-effect functions $\alpha_i(t)$, $i = 1, 2, \ldots, k$. In fact, they are not identifiable unless some constraint is imposed. If we do want to estimate these main-effect functions, the most commonly used constraint is

$$\sum_{i=1}^{k} n_i \alpha_i(t) = 0,$$

involving the k sample sizes. Under this constraint, it is easy to show that the unbiased estimators of the main-effect functions are

$$\hat{\alpha}_i(t) = \bar{y}_{i.}(t) - \bar{y}_{..}(t), \quad i = 1, 2, \ldots, k,$$

where

$$\bar{y}_{..}(t) = n^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}(t) = n^{-1} \sum_{i=1}^{k} n_i \bar{y}_{i.}(t)$$

is the usual sample grand mean function.

Let

$$SSH_n(t) = \sum_{i=1}^{k} n_i[\bar{y}_{i.}(t) - \bar{y}_{..}(t)]^2,$$

and

$$SSE_n(t) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} [y_{ij}(t) - \bar{y}_{i.}(t)]^2.$$

Theorem

Suppose Assumptions KS1 and KS2 hold. Then under the null hypothesis, we have

$$\int TSSH_n(t)dt \stackrel{d}{=} \sum_{r=1}^{m} \lambda_r A_r, \quad A_r \stackrel{i.i.d.}{\sim} \chi^2_{k-1},$$

$$\int TSSE_n(t)dt \stackrel{d}{=} \sum_{r=1}^{m} \lambda_r E_r, \quad E_r \stackrel{i.i.d.}{\sim} \chi^2_{n-k},$$

where A_r, E_r, $r = 1, 2, \ldots, m$ are independent of each other, and $\lambda_1, \ldots, \lambda_m$ are all the positive eigenvalues of $\gamma(s, t)$.

Theorem

Suppose Assumptions KS1, KS3, and KS4 hold. Then under the null hypothesis, as $n \to \infty$, we have

$$\int TSSH_n(t)dt \xrightarrow{d} \sum_{i=1}^{m} \lambda_r A_r, \quad A_r \overset{i.i.d.}{\sim} \chi^2_{k-1},$$

where $\lambda_1, \ldots, \lambda_m$ are all the positive eigenvalues of $\gamma(s, t)$.

The pointwise F-test is conducted at each $t \in T$ using the following pointwise F statistic:

$$F_n(t) = \frac{SSH_n(t)/(k-1)}{SSE_n(t)/(n-k)}.$$

From the classical linear model theory, it is easy to see that when the k samples are Gaussian, we have

$$F_n(t) \sim F_{k-1,n-k}, \quad t \in T.$$

L^2-Norm Based Test

$$T_n = \int_T SSH_n(t)dt = \sum_{i=1}^{k} n_i \int_T [\bar{y}_{i.}(t) - \bar{y}_{..}(t)]^2 dt.$$

we have or approximately have

$$T_n = \sum_{r=1}^{m} \lambda_r A_r, \quad A_r \overset{i.i.d.}{\sim} \chi^2_{k-1},$$

where $\lambda_r, r = 1, 2, \ldots, m$ are all the positive eigenvalues of $\gamma(s,t)$. It follows that we can approximate the null distribution of T_n by the Welch-Satterthwaite χ^2-approximation method. we obtain

$T_n \sim \beta\chi^2_{(k-1)\kappa}$ approximately, where $\beta = \dfrac{tr(\gamma \otimes 2)}{tr(\gamma)}, \quad \kappa = \dfrac{tr^2(\gamma)}{tr(\gamma \otimes 2)}.$

F Type Test

When the k samples are Gaussian, we can conduct an F-type test for the main-effect test. The F-type test statistic is defined as

$$F_n = \frac{\int_T SSH_n(t)dt}{(k-1)} \Big/ \frac{\int_T SSE_n(t)dt}{(n-k)}.$$

we have

$$F_n \stackrel{d}{=} \frac{\sum_{r=1}^m \lambda_r A_r}{(k-1)} \Big/ \frac{\sum_{r=1}^m \lambda_r E_r}{(n-k)},$$

where $A_r \stackrel{i.i.d.}{\sim} \chi^2_{k-1}$, $E_r \stackrel{i.i.d.}{\sim} \chi^2_{n-k}$ and they are all independent; $\lambda_1, \lambda_2, \ldots, \lambda_m$ are all the positive eigenvalues of $\gamma(s,t)$. We have

$$F_n \sim F_{(k-1)\hat{\kappa},(n-k)\hat{\kappa}} \text{ approximately,}$$

12

- Argentina,
- Brazil,
- Canada,
- Chile,
- Colombia,
- Costa Rica,
- Ecuador,
- Jamaica,
- Mexico,
- Peru,
- USA,
- Venezuela

33

- Austria, Belgium, Bosnia and Herzegovina,
- Bulgaria, Croatia, Cyprus,
- Czech Republic, Denmark, Finland,
- France, Germany, Greece,
- Hungary, Iceland, Ireland,
- Italy, Malta, Montenegro,
- Netherlands, Norway, Poland,
- Portugal, Romania, Russia,
- Serbia, Slovakia, Slovenia,
- Spain, Sweden, Switzerland,
- Turkey, Ukraine, United Kingdom

19

- Australia,
- Bangladesh,
- China,
- Hong Kong,
- India,
- Indonesia,
- Japan,
- Kazakhstan,
- Malaysia,
- Mongolia,
- New Zealand,
- Pakistan,
- Philippines,
- Singapore,
- South Korea,
- Sri Lanka,
- Taiwan,
- Thailand,
- Vietnam

31

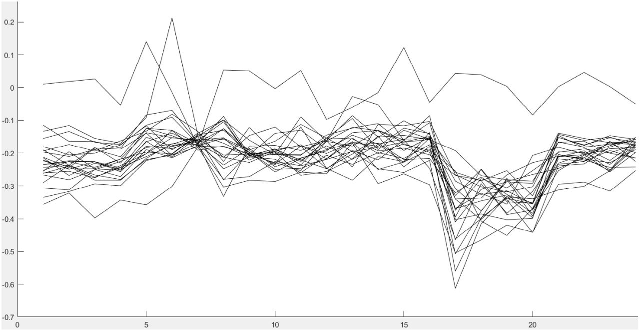


Figure: All Stock Exchange Data.

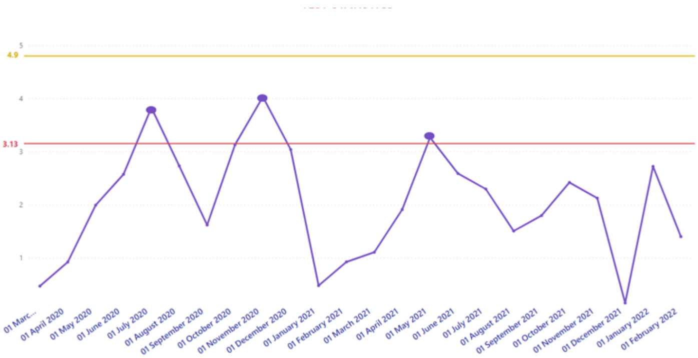


Figure: Example of Functional Data.

COVID-19 in the U.S.:

- July 2020 marked the first wave of economic disruptions caused by the pandemic, with the U.S. experiencing a sharp rise in cases while some regions in Europe and Asia showed signs of recovery.
- By December 2020, the rollout of vaccines in the U.S. created a sense of optimism, potentially driving different market behaviors compared to Europe and Asia.
- In May 2021, the U.S. had gained significant momentum in its economic recovery, particularly due to accelerated vaccination efforts, putting it ahead of Europe and Asia.

Economic Stimulus Packages

- The U.S. implemented massive stimulus packages to mitigate the economic effects of the pandemic:
- July 2020: Initial stimulus measures began to positively influence U.S. markets.
- December 2020: Discussions and approval of a second major stimulus package further boosted confidence.
- May 2021: Additional infrastructure and recovery plans under the Biden administration might have led to distinct market movements.
- In contrast, Europe and Asia may not have implemented stimulus packages as extensive or rapid as those in the U.S.

- Performance of Technology Stocks (e.g., Apple, Amazon, Tesla)
- Monetary Policy and Interest Rates
- Geopolitical and Trade Factors
- Investor Sentiment
- Sectoral Composition Differences

Three-way ANOVA

Consider the following three-way ANOVA model,

$$\begin{aligned} y_{ijkl} = \mu_0(t) + \alpha_i(t) + \beta_j(t) + \theta_k(t) + \alpha\beta_{ij}(t) + \alpha\theta_{ik}(t) + \beta\theta_{jk}(t) \\ + \alpha\beta\theta_{ijk}(t) + \epsilon_{ijkl}(t), \\ t \in T, i = 1, 2, ..., a; j = 1, 2, ..., b; k = 1, 2, ..., c; l = 1, 2, ..., n_{ijk} \end{aligned} \quad (1)$$

where $\mu_0(t)$ is the grand mean function, $\alpha_i(t)$, $\beta_j(t)$ and $\theta_k(t)$ are the i^{th}, j^{th} and k^{th} main-effect functions of factors A, B and C, respectively. $\alpha\beta_{ij}(t)$, $\alpha\theta_{ik}(t)$ and $\beta\theta_{jk}(t)$ are the interaction effect functions between the factors (A,B), (A,C) and (B,C). $\alpha\beta\theta_{ijk}(t)$ is the $(i, j, k)^{th}$ interaction effect function between the factors A, B and C. All these samples are assumed to be independent of each other and all the subject-effect functions

Consider the ANOVA model, For this model, we are interested in the following null hypotheses:

$$
\begin{aligned}
H_{0A} &: \alpha_i(t) = 0; i = 1, 2, ..., a, t \in T, \quad (2)\\
H_{0B} &: \beta_j(t) = 0; j = 1, 2, ..., b, t \in T,\\
H_{0C} &: \theta_i(t) = 0; k = 1, 2, ..., c, t \in T,\\
H_{0AB} &: \alpha\beta_{ij}(t) = 0; i = 1, 2, ..., a, j = 1, 2, ..., b, t \in T,\\
H_{0AC} &: \alpha\theta_{ik}(t) = 0; i = 1, 2, ..., a, k = 1, 2, ..., c, t \in T,\\
H_{0BC} &: \beta\theta_{jk}(t) = 0; j = 1, 2, ..., a, k = 1, 2, ..., c, t \in T,\\
H_{0ABC} &: \alpha\beta\theta_{ijk}(t) = 0; i = 1, 2, ..., a, j = 1, 2, ..., b, k = 1, 2, ..., c, t \in T,
\end{aligned}
$$

$$\hat{\eta_{ijk}}(t) = \bar{y_{ijk.}}(t) = n_{ijk}^{-1} \sum_{k=1}^{n_{ijk}} y_{ijkl}(t), \quad i = 1, 2, ..., a, j = 1, 2, ..., b, k = 1, 2, ..., c \tag{3}$$

Based on this, we can also estimate the common covariance function unbiasedly by the following pooled sample covariance function

$$\hat{\gamma(s, t)} = (n - abc)^{-1} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \sum_{l=1}^{n_{ijk}} \left[y_{ijkl}(s) - \bar{y_{ijk.}}(s) \right] \left[y_{ijkl}(t) - \bar{y_{ijk.}}(t) \right] \tag{4}$$

where $n = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} n_{ijk.}$. As all the *abc* cell samples are independent and have the common covariance function $\gamma(s, t)$

$$g_i = \sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}/n, \; h_j = \sum_{i=1}^{a}\sum_{k=1}^{c} n_{ijk}/n \text{ and } v_k = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ijk}/n$$

Let $\mathbf{g}$, $\mathbf{h}$ and $\mathbf{v}$ be the vector representations and $\mathbf{e}_{r,p}$ be p-dimensional unit vector whose rth component is 1 and others are zero, then we have

$$\eta_0(t) = \left[\mathbf{g}^T \otimes \mathbf{h}^T \otimes \mathbf{v}^T\right]\eta(t), \; \alpha_i(t) = \left[(\mathbf{e}_{i,a} - \mathbf{g})^T \otimes \mathbf{h}^T \otimes \mathbf{v}^T\right]\eta(t)$$

$$\beta_j(t) = \left[\mathbf{g}^T \otimes (\mathbf{e}_{i,b} - \mathbf{h})^T \otimes \mathbf{v}^T\right]\eta(t), \; \theta_k(t) = \left[\mathbf{g}^T \otimes \mathbf{h}^T \otimes (\mathbf{e}_{i,c} - \mathbf{v})^T\right]\eta(t)$$

$$\alpha\beta_{ij}(t) = \left[(\mathbf{e}_{i,a} - \mathbf{g})^T \otimes (\mathbf{e}_{i,b} - \mathbf{h})^T \otimes \mathbf{v}^T\right]\eta(t),$$

$$\alpha\theta_{ik}(t) = \left[(\mathbf{e}_{i,a} - \mathbf{g})^T \otimes \mathbf{h}^T \otimes (\mathbf{e}_{i,c} - \mathbf{v})^T\right]\eta(t)$$

$$\beta\theta_{jk}(t) = \left[\mathbf{g}^T \otimes (\mathbf{e}_{i,b} - \mathbf{h})^T \otimes (\mathbf{e}_{i,c} - \mathbf{v})^T\right]\eta(t)$$

and

$$\eta_0(t) = \left[\mathbf{g}^T \otimes \mathbf{h}^T \otimes \mathbf{v}^T\right]\eta(t) = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n_{ijk}} y_{ijkl}(t)$$

$$A_a = [\mathbf{I}_a - \mathbf{1}_a\mathbf{g}^T] \otimes \mathbf{h}^T \otimes \mathbf{v}^T, \; A_b = \mathbf{g}^T \otimes [\mathbf{I}_b - \mathbf{1}_b\mathbf{h}^T] \otimes \mathbf{v}^T$$
$$A_c = \mathbf{g}^T \otimes \mathbf{h}^T \otimes [\mathbf{I}_c - \mathbf{1}_c\mathbf{v}^T], \; A_{ab} = [\mathbf{I}_a - \mathbf{1}_a\mathbf{g}^T] \otimes [\mathbf{I}_b - \mathbf{1}_b\mathbf{h}^T] \otimes \mathbf{v}^T$$
$$A_{ac} = [\mathbf{I}_a - \mathbf{1}_a\mathbf{g}^T] \otimes \mathbf{h}^T \otimes [\mathbf{I}_c - \mathbf{1}_c\mathbf{v}^T], A_{bc} = \mathbf{g}^T \otimes [\mathbf{I}_b - \mathbf{1}_b\mathbf{h}^T] \otimes [\mathbf{I}_c - \mathbf{1}_c\mathbf{v}^T$$

and

$$A_{abc} = [\mathbf{I}_a - \mathbf{1}_a\mathbf{g}^T] \otimes [\mathbf{I}_b - \mathbf{1}_b\mathbf{h}^T] \otimes [\mathbf{I}_c - \mathbf{1}_c\mathbf{v}^T]$$

It can be noticed that $A_a, A_b, A_c, A_{ab}, A_{ac}, A_{bc}$ and A_{abc} are not full-rank matrices and have $(a-1), (b-1), (c-1), (a-1)(b-1), (a-1)(c-1), (b-1)(c-1)$ and $(a-1)(b-1)(c-1)$ ranks respectively.

$$\hat{\alpha}_i(t) = \left[(\mathbf{e}_{i,a} - \mathbf{g})^T \otimes \mathbf{h}^T \otimes \mathbf{v}^T\right]\eta(t) = A_a\hat{\eta}(t)$$

And define the following hypothesis matrices

$$\mathbf{H}_a = (\mathbf{I}_{a-1}, -\mathbf{1}_{a-1}),\ \mathbf{H}_b = (\mathbf{I}_{b-1}, -\mathbf{1}_{b-1}),\ \mathbf{H}_c = (\mathbf{I}_{c-1}, -\mathbf{1}_{c-}$$
$$\mathbf{H}_{ab} = \mathbf{H}_a \otimes \mathbf{H}_b,\ \mathbf{H}_{ac} = \mathbf{H}_a \otimes \mathbf{H}_c,\ \mathbf{H}_{bc} = \mathbf{H}_b \otimes \mathbf{H}_c,\ \text{and } \mathbf{H}_{abc} = \mathbf{H}_a \otimes \mathbf{H}_b \otimes$$

and

$$H_{0A} : \alpha_i(t) = 0; i = 1, 2, ..., a, t \in T = \mathbf{H}_a \alpha(t) = 0$$

$$\mathbf{S}_a = \mathbf{H}_a\mathbf{A}_a,\ \mathbf{S}_b = \mathbf{H}_b\mathbf{A}_b,\ \mathbf{S}_c = \mathbf{H}_c\mathbf{A}_c,\ \mathbf{S}_{ab} = H_{ab}\mathbf{A}_{ab},\ \mathbf{S}_{ac} = \mathbf{H}_{ac}\mathbf{A}_{ac},$$
$$\mathbf{S}_{bc} = \mathbf{H}_{bc}\mathbf{A}_{bc},\ \text{and } \mathbf{S}_{abc} = \mathbf{H}_{abc}\mathbf{A}_{abc},$$

where **S** becomes a full-rank matrices with q. Therefore, the sum of squares of the corresponding null hypothesis and the sum of squares are obtained as

$$SSH_n(t) = [\mathbf{S}\hat{\eta}(t)]^T(\mathbf{SDS}^T)^{-1}[\mathbf{S}\hat{\eta}(t)]$$

and

$$SSE_n(t) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n_{ijk}}\left[y_{ijkl}(t) - \bar{y_{ijk\cdot}}(t)\right]^2 = (n - abc)\hat{\gamma}(t,t).$$

Using pointwise testing method, we define the following test statistics for each null hypothesis as,

$$F_n(t) = \frac{SSH_n(t)/q}{SSE_n(t)/(n - abc)}$$

It can be noticed that the F statistic is obtained by using Pointwise test methodology under the normality assumption. If the normality assumption is not satisfied, $\mathcal{L}^2$-norm-based test, Welch-Satterthwaite χ^2 approximation method or F-Type Test can be used. .

In this application, the scale is applied to 43 health employees in Turkey between January 2021 and December 2021 for monthly. The main purpose of the project is to determine the anxiety and depression levels of health employees during Covid-19 pandemic. In this paper, we use only the anxiety point of the employees. Figure shows the data in functional form.

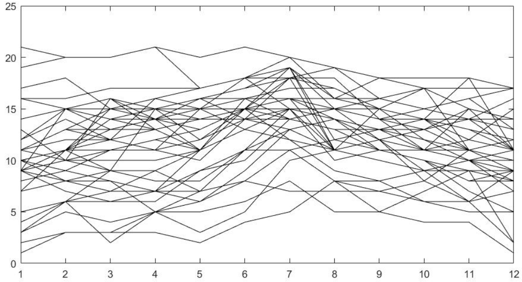
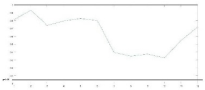


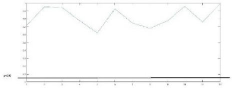
Figure: Example of Functional Data.

In this data, there are three treatments: the gender (23 female, 20 male), the profession (20 doctor, 23 nurse) and the age (13 25, 25-40 and 14 40+). There are two levels of gender, two levels of the profession and three level of the age also this data have 12 different time of measurement. Our motivation is to determine is there any significant difference between the treatment in different time points.

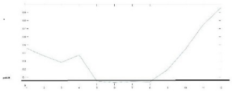
(Gender)



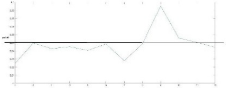
(Profession)



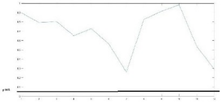
(Age)



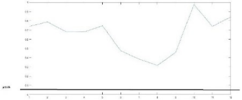
(Gender & Profession)



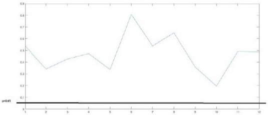
(Gender & Age)



(Profession & Age)



(Gender & Profession & Age)



- Schmidt SJ, Barblan LP, Lory I, Landolt MA (2021). ”Age-related effects of the COVID-19 pandemic on mental health of children and adolescents.” *European Journal of Psychotraumatology*, Taylor Francis.
- Gamsızkan Z, Sungur MA, Smith R (2021). ”How do older age, gender, and risk groups affect protective behaviours and mental health in the COVID-19 pandemic?” *International Journal of Mental Health*, Wiley Online Library.
- Steinegger B, Arola-Fernández L, Chen X (2022). ”Behavioural response to heterogeneous severity of COVID-19 explains temporal variation of cases among different age groups.” *Proceedings of the Royal Society B*, Royal Society Publishing.
- Atchison C, Bowman LR, Vrinten C, Redd R (2021). ”Early perceptions and behavioural responses during the COVID-19 pandemic: a cross-sectional survey of UK adults.” *BMJ Open*, BMJ Publishing.
- Balkhi F, Nasir A, Zehra A, Riaz R (2020). ”Psychological and behavioral response to the coronavirus (COVID-19) pandemic.” *Cureus*, NCBI.

Othe Functional Data Analysis

- Principal Component Analysis
- Canonical Correlation
- Clustering Analysis
- Discriminant Analysis
- Regression Models and Logistic Regression

Thank you for listening :)

References

Benhenni K, Ferraty F, Rachdi M, Vieu P (2007). Local smoothing regression with functional data. Computational Statistics, 22, 353–369

Berrendero JR, Justel A, Svarc M (2011). Principal components for multivariate functional data. Computational Statistics and Data Analysis, 55, 2619–2634

Boente G, Fraiman R (2000). Kernel-based functional principal components. Ann Stat, 20, 655–674

Cai T, Hall P (2006). Prediction in functional linear regression. Ann Stat, 34, 2159–2179

Chiou JM, Müller HG (2007). Diagnostics for functional regression via residual processes. Comput Stat Data Anal, 15, 4849–4863

Cuevas A (2014) A partial overview of the theory of statistics with

Cuevas A, FebreroM, Fraiman R (2004). An anova test for functional data. Comput Stat Data Anal, 47, 111– 122

Górecki T, Krzy´skoM,Waszak Ł (2014). Functional discriminant coordinates. Commun Stat Theory Methods, 43, 1013–1025

James, G. M (2002). Generalized linear models with functional predictors. Journal of the Royal Statistical Society, Series B (Statistical Methodology), 64, 411–432

James GM, Hastie TJ (2001). Functional linear discriminant analysis for irregularly sampled curves. J R Stat Soc Ser B (Stat Methodol), 63, 533–550

Krzy´sko M, Waszak Ł (2013). Canonical correlation analysis for functional data. Biometr Lett, 50, 95–105

Leurgans SE, Moyeed RA, Silverman BW(1993). Canonical correlation analysis when the data are curves. J R Stat Soc Ser B (Stat Methodol), 55, 725–740

Preda C, Saporta G, Lévéder C (2007) PLS classification of functional data. Comput Stat 22:223–235

Ramsay, J.O., Dalzell, C.J. (1991). Some tools for functional data analysis (with discussion). J. Roy. Statist. Soc. B 52, 539–572.

Ramsay, J.O., Silverman, B.W., 1997. Functional Data Analysis.. Springer, New York.

Ramsay, J.O., Silverman, B.W., 2002. AppliedFunctional Data Analysis. Springer, New York.

Ratcliffe S.J., Heller G.Z. and Leader L.R. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression, Statistics in Medicine, 21, 1115–1127

Tokushige S, Yadohisa H, Inada K (2007). Crisp and fuzzy k-means clustering algorithms for multivariate functional data. Comput Stat, 22, 1–16

Yamamoto M, Terada Y (2014). Functional factorial K-means analysis. Comput Stat Data Anal, 79, 33–148

Zhang JT (2013) Analysis of variance for functional data. Chapman and Hall, London

The End

Questions? Comments?