

# **GIỚI THIỆU**

## **THỰC TẬP ỨNG DỤNG THỐNG KÊ**

### **TRONG NGHIÊN CỨU SINH HỌC**

#### **CSH139 – SHH681 – BTE10019**

**Nguyễn Thị Gia Hằng**

**Đặng Lê Anh Tuấn**

**Nguyễn Trần Thụy Thanh Mai**

**TLTK: Thực tập Ứng dụng thống kê trong nghiên cứu Sinh học**

**Chuyên ngành: Sinh thái – Tài nguyên sinh vật**

**Bộ môn Sinh thái – Sinh học tiến hóa**

**Khoa Sinh học – Công nghệ sinh học**

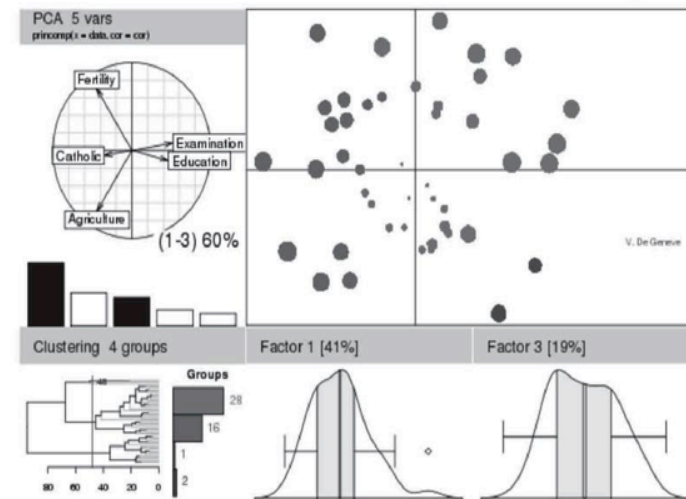
**Trường Đại học Khoa học Tự nhiên – ĐHQG TpHCM**

ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐẶNG LÊ ANH TUẤN  
NGUYỄN THỊ GIA HẰNG  
NGUYỄN TRẦN THỤY THANH MAI



## THỰC TẬP ỨNG DỤNG THỐNG KÊ TRONG NGHIÊN CỨU SINH HỌC



 NHÀ XUẤT BẢN  
ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

ĐẶNG LÊ ANH TUẤN  
NGUYỄN THỊ GIA HẰNG  
NGUYỄN TRẦN THỤY THANH MAI

THỰC TẬP ỨNG DỤNG THỐNG KÊ  
TRONG NGHIÊN CỨU SINH HỌC

NXB ĐHQG HCM



# ĐỊA ĐIỂM HỌC THỰC HÀNH và MỤC TIÊU BUỔI HỌC

**Địa điểm:** Trung tâm Tin học Trường ĐH KHTN,  
137E Nguyễn Chí Thanh, Quận 5, TpHCM.

**Sau buổi giới thiệu này, sinh viên (SV):**

- biết được các Thầy Cô tham gia giảng dạy và vai trò chính của từng Thầy Cô
- nắm được nội dung chính của môn học TT UDTKTNCSSH
- biết được một số ứng dụng của R

# ĐÁNH GIÁ

- **Thi cuối kì (CK):** thực hành trên máy tính để điền kết quả vào phiếu trả lời (trắc nghiệm điền khuyết): 10/10.
- **Mã SHH681 và CSH139 (3 tín chỉ = 2LT + 1TT):** điểm thực tập (điểm giữa kì, 40%) cộng chung với điểm lý thuyết (60%, Mã BTE10014, Thầy Long phụ trách) = 1 cột điểm. Điều kiện qua môn học: mỗi cột điểm GK và CK đạt ít nhất 5 điểm.
- **Mã BTE10019 (1 tín chỉ TT):** điểm thực tập được tính riêng, 1 cột CK, **tách biệt** với BTE10014 (2 tín chỉ LT).
- Bài tập trên lớp + về nhà

**Phụ trách môn Lý thuyết UDTKTNC SH** (CSH139, SHH681 và BTE10014): Thầy Nguyễn Mỹ Phi Long – PTN Vi sinh

**Phụ trách môn Thực tập UDTKTNC SH** (CSH139, SHH681 và BTE10019): Cô Nguyễn Thị Gia Hằng, Thầy Đặng Lê Anh Tuấn, Cô Nguyễn Trần Thuy Thanh Mai – Bộ môn Sinh thái – Sinh học tiến hóa (Chuyên ngành Sinh thái – Tài nguyên sinh vật).

**Vai trò chính:**

- Cô Hằng: Trưởng thực tập, phụ trách chung
- Thầy Tuấn: phụ trách lý thuyết thực tập và giải thích kết quả
- Cô Mai: phụ trách kỹ thuật

## CÁC BƯỚC THỰC HIỆN 1 NGHIÊN CỨU

- 1/ Ý tưởng, tổng quan tài liệu
- 2/ **Mục tiêu/ câu hỏi nghiên cứu**
- 3/ **Thiết kế thí nghiệm** (Experiment design)
- 4/ **Thí nghiệm thử nghiệm** (pilot study) và **ước tính cỡ mẫu**
- 5/ Thí nghiệm và thu thập dữ liệu
- 6/ Biên tập và **phân tích dữ liệu**
- 7/ **Giải thích và báo cáo kết quả phân tích**

1. **Ngày 02/10:** giới thiệu R, các bước nghiên cứu, nhập và biên tập số liệu (Chương 1-3)
2. **Ngày 09/10:** xác định giá trị ngoại lệ và thống kê mô tả bằng thông số (Chương 4);
3. **Ngày 16/10:** thống kê mô tả bằng hình ảnh (Chương 5);
4. **Ngày 23/10:** t-test độc lập (Chương 6, mục 6.1);
5. **Ngày 30/10:** t-test bắt cặp (Chương 6, mục 6.2);
6. **Ngày 06/11:** phân tích phương sai một nhân tố (Chương 7, mục 7.1);
7. **Ngày 13/11:** phân tích phương sai hai nhân tố (Chương 7, mục 7.2);
8. **Ngày 20/11: Nghỉ lễ Nhà giáo VN**
9. **Ngày 27/11:** hồi qui tuyến tính đơn biến (Chương 8);
10. **Ngày 04/12:** Chi-squared test ( $\chi^2$ -test, Chương 9) + Giải đáp thắc mắc
11. **Ngày 11/12:** SV ôn tập tại nhà
12. **Ngày 18/12: Thi cuối kì**

- Tựa bài
- Mục tiêu bài học
- Các điều kiện cần thỏa để sử dụng một kiểm định thống kê
- Các bước thực hiện
- Câu lệnh và ý nghĩa các câu lệnh
- Hướng dẫn đọc và giải thích kết quả phân tích
- Hướng dẫn chọn, ghi và ý nghĩa của kết quả
- Bài tập về nhà



## XEM VIDEO

**link: [https://  
www.youtube.com/watch?  
v=XcBLEVknqvY](https://www.youtube.com/watch?v=XcBLEVknqvY)**

THÀNH VIÊN

NỘI DUNG

R

Q & A



## The R Project for Statistical Computing

About R

[What is R?](#)

[Contributors](#)

[Screenshots](#)

[What's new?](#)

Download,  
Packages

[CRAN](#)

R Project

[Foundation](#)

[Members & Donors](#)

[Mailing Lists](#)

[Bug Tracking](#)

[Developer Page](#)

[Conferences](#)

[Search](#)

Documentation

[Manuals](#)

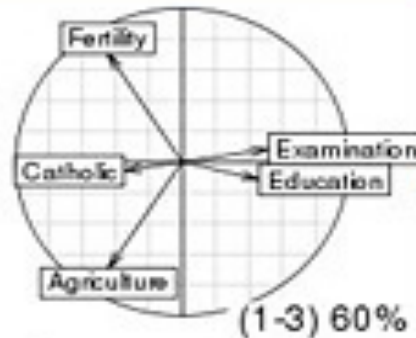
[FAQs](#)

[The R Journal](#)

[Wiki](#)

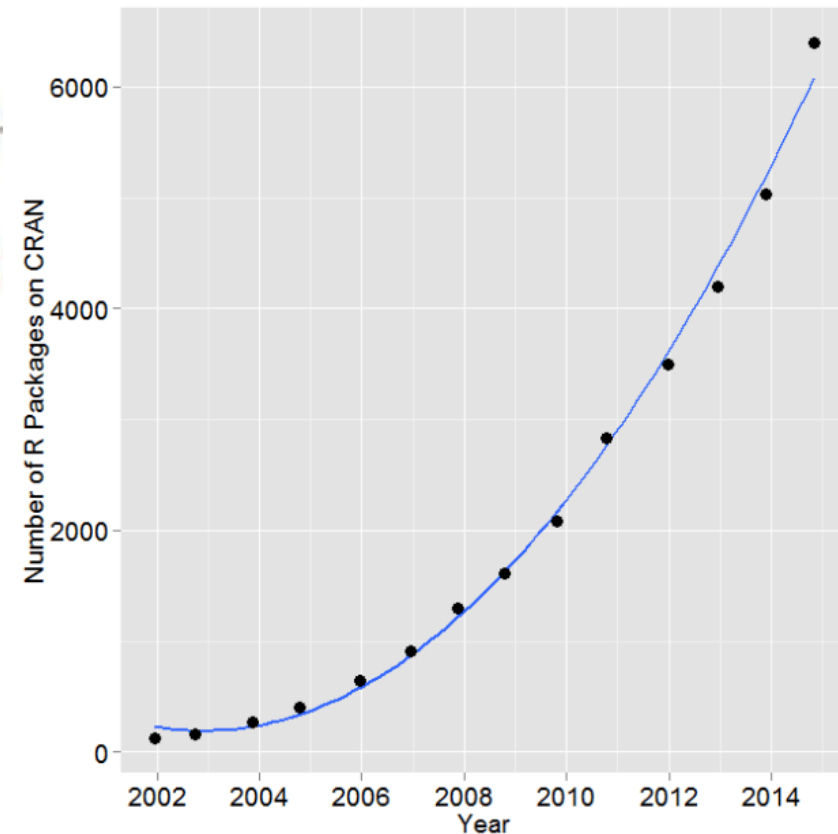
[Books](#)

PCA 5 vars  
`princomp(x = data, cor = cor)`



### Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).



THÀNH VIÊN

NỘI DUNG

R

Q & A

**Q & A**

# BONUS

# MẪU FILE NHẬP SỐ LIỆU KHÔNG BẮT CẤP

CHÚ GIẢI: subject = đơn vị thí nghiệm; vegtype = kiểu sinh cảnh; season = mùa; BA = diện tích lá đước đôi (cm<sup>2</sup>); BDW = trọng lượng khô lá đước đôi (g); địa điểm RNM Cần Giờ

<b>subject</b>	<b>vegtype</b>	<b>season</b>	<b>BA</b>	<b>BDW</b>
1	F	13wet	23.13	0.19
2	F	13wet	34.55	0.31
3	F	13wet	23.72	0.20
4	O	13wet	18.45	0.20
5	O	13wet	24.08	0.35
6	O	13wet	33.92	0.50
7	F	14dry	39.36	0.47
8	F	14dry	21.46	0.23
9	F	14dry	64.56	0.71
10	O	13wet	31.57	0.39
11	O	13wet	21.67	0.25
12	O	13wet	17.58	0.22

# MẪU FILE NHẬP SỐ LIỆU BẮT CẬP

**CHÚ GIẢI:** subject = đơn vị thí nghiệm; time = trước (1before) và sau (2after) khi thí nghiệm; weight = cân nặng cơ thể (kg); địa điểm BV Chợ Rẫy

<b>subject</b>	<b>time</b>	<b>weight</b>
1	1before	83.8
2	1before	83.3
3	1before	86
4	1before	82.5
5	1before	86.7
6	1before	79.6
1	2after	95.2
2	2after	94.3
3	2after	91.5
4	2after	91.9
5	2after	100.3
6	2after	76.7

## VARIABLE: vai trò

$$Y = aX + b$$

**X = biến độc lập** (independent/ explanatory variable/ predictor):

- Dễ xác định (đo, phân hạng): đường kính cây, độ tuổi, giới tính
- Được nhà nghiên cứu “**cho là**” có ảnh hưởng lên biến phụ thuộc Y nên được dùng để giải thích cho sự thay đổi của biến Y

**Y = biến phụ thuộc** (dependent/ explained/ predicted/ response/ outcome variable):

- Khó xác định (đo, phân hạng, ...) hơn X: chiều cao cây, cholesterol
- Được nhà nghiên cứu “**cho là**” chịu ảnh hưởng của biến độc lập X

## VARIABLE: tính chất (cho cả X và Y)

**Biến định lượng** (quantitative/ numeric variable): là biến có thể cân, đo, đong, đếm được (VD: độ tuổi, độ cao, chiều dài, trọng lượng, số cá thể, số khuẩn lạc, ...)

**Biến liên tục** (continuous variable):

biến có các giá trị giữa 2 số

**Biến rời rạc** (discrete variable):

biến chỉ có các giá trị số nguyên



## VARIABLE: tính chất (cho cả X và Y)

**Biến định tính** (qualitative/ categorical variable): là biến không đo được mà có tính phân hạng hay danh tính (VD: có/ không, mùa mưa/ mùa khô, màu sắc, giống vi sinh vật, giống cây trồng, ...)

**Biến tên gọi/ định tính không phân hạng** (nominal variable): biến có các giá trị (levels) là tên gọi và tên gọi này không có trật tự hoặc trật tự của tên gọi không có ý nghĩa (VD: đỏ, đen, trắng; lúa mì, lúa nước, lúa mạch).

**Biến định tính phân hạng** (ordinal/ rank variable): biến định tính có các giá trị (levels) có tính phân hạng (VD: nhất, nhì, ba; giỏi, khá, trung bình).

# GIẢ THIẾT, SAI LẦM I (alpha), II (beta), POWER và Effect size

$H_0$ : Không có sự khác biệt/ ảnh hưởng/ tương quan/ liên hệ

$H_A$ : Có sự khác biệt / ảnh hưởng/ tương quan/ liên hệ

		Reality	
		$H_0$ False	$H_0$ True
Test	Reject $H_0$	Correct rejection $H_0$ ✓ = Power = $1 - \beta$	✗ Type I error = $\alpha$
	Accept $H_0$	✗ Type II error	Correct acceptance ✓ of $H_0$

## GIẢ THIẾT, SAI LẦM I (alpha), II (beta), POWER và Effect size

**Alpha** (significant level): mức ý nghĩa = sai lầm loại 1: xác suất để kết quả kiểm định thống kê BÁC BỎ  $H_0$  (nhận  $H_A$ ) khi trên thực tế  $H_0$  đúng. Vậy alpha = xác suất BÁC BỎ sai  $H_0$ .

**Beta** = sai lầm loại 2: xác suất để kết quả kiểm định thống kê NHẬN  $H_0$  (bác bỏ  $H_A$ ) khi trên thực tế  $H_0$  sai. Vậy beta = xác suất NHẬN sai  $H_0$ .

**Power** =  $1 - \text{beta}$ : xác suất để kết quả kiểm định thống kê BÁC BỎ  $H_0$  (nhận  $H_A$ ) khi trên thực tế  $H_0$  sai. Vậy power = xác suất BÁC BỎ đúng  $H_0$ .

**Effect size**: định lượng cường độ (strength) của một tác động/ ảnh hưởng/ tương quan/ liên hệ

## TRUE vs. PSEUDO-REPLICATION

(True) replications, also known as experimental/ observation units, are defined as **independent** applications of the same treatment

=> Right conclusion from statistic analysis.

Pseudo-replication occurs when **experimental/ observation units are not statistically independent, but treated as if they are.**

=> Wrong conclusion from statistic analysis.

**VD1: Ảnh hưởng của phân đạm lên lúa.**

**VD2: Nuôi chuột**

**VD3: Thí nghiệm lá gan chuột**

- 1/ Ý tưởng, tổng quan tài liệu
- 2/ **Mục tiêu/ câu hỏi nghiên cứu**
- 3/ **Thiết kế thí nghiệm** (Experiment design)
- 4/ **Thí nghiệm thử nghiệm** (pilot study) và **ước tính cỡ mẫu**
- 5/ Thí nghiệm và thu thập dữ liệu
- 6/ Biên tập và **phân tích dữ liệu**
- 7/ **Giải thích và báo cáo kết quả phân tích**

## SO SÁNH TRUNG BÌNH CỦA 2 NHÓM KHÔNG BẮT CẶP (độc lập)

$$n_i = \frac{2 \left( Z_{\frac{\alpha}{2}} + Z_{\beta} \right)^2 \sigma^2}{(\mu_1 - \mu_2)^2}$$

Các thông tin cần biết để ước tính cỡ mẫu so sánh 2 nhóm:

- Effect size ( $\mu_1 - \mu_2$ ) và  $\sigma$  có được từ nghiên cứu trước/ pilot study
- Alpha
- Beta (hay  $1 - \beta = \text{Power}$ )
- **Kết quả: Tổng số mẫu cần thiết:  $2 \times n_i$**

## SO SÁNH TRUNG BÌNH CỦA 2 NHÓM BẮT CẶP

$$n_i = \frac{\left(Z_{\frac{\alpha}{2}} + Z_{\beta}\right)^2 \sigma^2}{2(\mu_1 - \mu_2)^2}$$

Các thông tin cần biết để ước tính cỡ mẫu so sánh 2 nhóm:

- Effect size ( $\mu_1 - \mu_2$ ) và  $\sigma$  có được từ nghiên cứu trước/ pilot study
- Alpha
- Beta (hay  $1 - \beta = \text{Power}$ )
- **Kết quả: Tổng số mẫu cần thiết:  $2 \times n_i$**

# BẢNG TRA GIÁ TRỊ $\left(Z_{\frac{\alpha}{2}} + Z_{\beta}\right)^2$

Two-sided	One-sided $\beta$				
$\alpha$	0.05	0.10	0.15	0.20	0.50
0.001	24.358	20.904	18.723	17.075	10.828
0.005	19.819	16.717	14.772	13.313	7.879
0.01	17.814	14.879	13.048	11.679	6.635
0.02	15.770	13.017	11.308	10.036	5.412
0.05	12.995	10.507	8.978	7.849	3.841
0.1	10.822	8.564	7.189	6.183	2.706
0.2	8.564	6.569	5.373	4.508	1.642
0.4	6.183	4.508	3.527	2.833	0.708



# R PACKAGES HỖ TRỢ ƯỚC TÍNH CỖ MẪU

1/ pwr và pwr2: ước lượng cỡ mẫu cho 1 tỉ lệ, so sánh 2 hay nhiều tỉ lệ, 1 trung bình, so sánh trung bình của 2 hay nhiều nhóm, hệ số tương quan/ hồi quy tuyến tính đơn biến,

2/ samplesize

3/ powerAnalysis

4/ epicalc

# CÁC KIỂM ĐỊNH THAM SỐ CƠ BẢN

## A/ Mục tiêu so sánh

- 1/ 2 nhóm: t.test (số liệu độc lập và bắt cặp)
- 2/ nhiều nhóm 1 nhân tố: anova one-way
- 3/ nhiều nhóm hai nhân tố: anova two-way
- 4/ Hai hay nhiều tỉ lệ: Chi-squared test

## B/ Mục tiêu tương quan/ hồi quy tuyến tính

- 1/ Tương quan đơn biến: hệ số tương quan Pearson
- 2/ Hồi quy tuyến tính đơn/ đa biến

## T.TEST cho 2 nhóm không bắt cặp (độc lập)

**Mục tiêu:** so sánh giá trị trung bình của 2 nhóm không bắt cặp (mỗi đối tượng thí nghiệm chỉ xuất hiện ở 1 trong 2 level của factor X). Phương trình toán tổng quát: **1 biến Y định lượng** theo **1 biến X định tính có 2 levels** (factor).

**VD:** so sánh chiều cao trung bình của nam và nữ trong lớp học.

Y = chiều cao; X = giới tính gồm 2 levels nam và nữ.

**Giả thiết:**

$H_0$ : Trung bình chiều cao của nam KHÔNG khác chiều cao của nữ (Giới tính KHÔNG có ảnh hưởng lên chiều cao)

$H_A$ : Trung bình chiều cao của nam KHÁC chiều cao của nữ  
(Giới tính CÓ ảnh hưởng lên chiều cao)

## T.TEST cho 2 nhóm không bắt cặp (độc lập)

**Giả định cần kiểm tra và thỏa** trước khi thực hiện t.test (validity of statistical test): số liệu của mỗi nhóm tuân theo phân phối chuẩn.

### **Trình tự thực hiện:**

- Nhập và lưu số liệu trong excel
- Chuyển số liệu vào phần mềm và tìm hiểu số liệu
- Xác định các thông số thống kê và kiểm định phân phối chuẩn
- Vẽ đồ thị hộp/ cột (phỏng đoán các giả định và kết quả phân tích thống kê)
- Kiểm định phương sai (t.test không bắt cặp có phương sai bằng nhau hay không bằng nhau)
- Kiểm định t.test không bắt cặp
- Kết luận

## T.TEST cho 2 nhóm không bắt cặp (độc lập)

**Kết luận:**

**Cách ghi kết quả so sánh phương sai:**

Kết quả **kiểm định phương sai** cho thấy phương sai chiều cao của nam (SD = ?) **có sự khác biệt** rõ rệt với nữ (SD = ?) ( $F = ?$ ; **P-value** = ? < **alpha** = ?, two-sided) với 95% C.I. = [giới hạn dưới; giới hạn trên].

**Cách ghi kết quả t.test không bắt cặp:**

Kết quả **phân tích t.test** hai nhóm không bắt cặp cho thấy chiều cao của nam (mean = ?, SD = ?) **có sự khác biệt** rõ rệt với nữ (mean = ?, SD = ?) ( $df = ?$ ;  $t = ?$ ; **P-value** = ? < **alpha** = ?, two-sided) với 95% C.I. = [giới hạn dưới; giới hạn trên].

## T.TEST cho 2 nhóm bất cặp

**Mục tiêu:** so sánh giá trị trung bình của 2 nhóm bất cặp (mỗi đối tượng thí nghiệm xuất hiện trong cả 2 level của factor X). Phương trình toán tổng quát: **1 biến Y định lượng** theo **1 biến X định tính có 2 levels** (factor).

**VD:** so sánh cân nặng của người nữ bị biếng ăn sau 3 tháng áp dụng 1 phương pháp điều trị.

Y = cân nặng cơ thể; X = thời gian gồm 2 levels trước và sau khi điều trị.

**Giả thiết:**

$H_0$ : Trung bình cân nặng trước khi điều trị KHÔNG khác sau khi điều trị (Phương pháp điều trị KHÔNG có ảnh hưởng/ hiệu quả)

$H_A$ : Trung bình cân nặng trước khi điều trị KHÁC sau khi điều trị (Phương pháp điều trị CÓ ảnh hưởng)

=>  $H_A$ : Trung bình cân nặng sau khi điều trị CAO hơn trước khi điều trị (Phương pháp điều trị CÓ hiệu quả)

## T.TEST cho 2 nhóm bất cặp

**Giả định cần kiểm tra và thỏa** trước khi thực hiện t.test (validity of statistical test): hiệu số của hai nhóm tuân theo phân phối chuẩn.

**Trình tự thực hiện:**

- Nhập và lưu số liệu trong excel
- Chuyển số liệu vào phần mềm và tìm hiểu số liệu
- Xác định các thông số thống kê và kiểm định phân phối chuẩn
- Vẽ đồ thị hộp/ cột (phỏng đoán các giả định và kết quả phân tích thống kê)
- Kiểm định t.test bất cặp
- Kết luận

## T.TEST cho 2 nhóm bắt cặp

**Kết luận:**

**Cách ghi kết quả t.test bắt cặp:**

Kết quả phân tích t.test hai nhóm bắt cặp cho thấy **có sự thay đổi trọng lượng cơ thể** sau khi điều trị (mean = ?, SD = ?) so với trước khi điều trị (mean = ?, SD = ?) [df = ?, t = ?; **P-value = ? < alpha = ?**, two tailed]. Trọng lượng cơ thể tăng ?kg (mean **wa** - mean **wb**) sau khi điều trị với 95% C.I. = [giới hạn dưới; giới hạn trên].



## Phân tích phương sai (Anova)

**Mục tiêu:** so sánh giá trị trung bình của nhiều nhóm không bắt cặp (between group anova) hay bắt cặp (within group anova). Phương trình toán tổng quát: **1 biến Y định lượng** theo **1 biến X định tính có nhiều levels** (factor).

**VD** cho between group one-way anova: so sánh diện tích phiến lá được đo giữa 3 khu vực thu mẫu B, C và D ở Cần Giờ.

Y = diện tích phiến lá; X = khu vực thu mẫu gồm 3 levels B, C và D.

### Giả thiết:

$H_0$ : Trung bình diện tích phiến lá được đo KHÔNG khác nhau giữa 3 khu vực B, C và D (Khu vực thu mẫu KHÔNG có ảnh hưởng lên diện tích phiến lá)

$H_A$ : Trung bình diện tích phiến lá được đo KHÁC nhau giữa 3 khu vực B, C và D (Khu vực thu mẫu CÓ ảnh hưởng lên diện tích phiến lá)

## Phân tích phương sai (Anova)

**Giả định cần kiểm tra và thỏa** trước khi thực hiện t.test (validity of statistical test):

- Không có sai sót ngẫu nhiên trong đo đạc (phụ thuộc vào thiết kế thí nghiệm và phương pháp đo)
- Các giá trị phần dư liên tục của mô hình phân tích không tương quan (độc lập) với nhau
- Phần dư của các nhóm có phương sai giống nhau
- Phần dư của mô hình phân tích theo phân phối chuẩn
- Phần dư của mô hình phân tích có trung bình = 0

# Phân tích phương sai (Anova)

## Trình tự thực hiện:

- Nhập và lưu số liệu trong Excel
- Chuyển số liệu vào phần mềm và tìm hiểu số liệu
- Xác định các thông số thống kê
- Vẽ đồ thị hộp (boxplot)
- Phân tích phương sai một nhân tố
  - Đặt mô hình phân tích phương sai một nhân tố và xem kết quả mô hình
  - Kiểm tra các điều kiện cần thỏa của mô hình Anova một nhân tố
  - So sánh từng cặp giá trị trung bình bằng phương pháp Tukey
  - Kết luận

## Phân tích phương sai (Anova)

### Kết luận

#### Cách ghi kết quả *ANOVA one-way*:

Kết quả phân tích phương sai một nhân tố cho thấy trung bình diện tích phiến lá BA giữa ba khu vực có sự **khác biệt rõ rệt về mặt thống kê** [ $F$  (df transect, df residuals) = ?; **P-value** = ? < **alpha** = ?]. So sánh cặp theo phân tích Tukey HSD cho thấy trung bình biến BA tại đường cắt ? (mean = ?, SD = ?) có sự **khác biệt rõ rệt** với đường cắt ? (mean = ?, SD = ?) với 95% C.I. [giới hạn dưới; giới hạn trên] và **P-value** = ? < **alpha** = ?.

## Hồi quy tuyến tính đơn biến

**Mục tiêu:** lập phương trình hồi quy tuyến tính giữa **1 biến Y định lượng** (phức tạp, khó đo) và **1 biến X định lượng** (đơn giản, dễ đo hơn) để **hỗ trợ nội suy giá trị** của biến **Y từ giá trị** của biến **X**.

**VD:** Ước tính hàm lượng Cholesterol trong máu theo độ tuổi của người nam.

Y = hàm lượng Cholesterol trong máu; X = độ tuổi của nam giới.

### Giả thiết:

$H_0$ : Độ tuổi KHÔNG giải đoán/ ước tính được hàm lượng Cholesterol trong máu của nam giới (Độ tuổi KHÔNG có ảnh hưởng lên hàm lượng Cholesterol trong máu của nam giới)

$H_A$ : Độ tuổi giải đoán/ ước tính được hàm lượng Cholesterol trong máu của nam giới (Độ tuổi CÓ ảnh hưởng lên hàm lượng Cholesterol trong máu của nam giới)

# Hồi quy tuyến tính đơn biến

**Giả định cần kiểm tra và thỏa** trước khi thực hiện t.test (validity of statistical test):

- Không có sai sót ngẫu nhiên trong đo đạc (phụ thuộc vào thiết kế thí nghiệm và phương pháp đo)
- Biến độc lập X và biến phụ thuộc Y có mối quan hệ tuyến tính.
- Các giá trị phần dư liên tục của mô hình phân tích không tương quan (độc lập) với nhau
- Phần dư của các nhóm có phương sai giống nhau
- Phần dư của mô hình phân tích theo phân phối chuẩn
- Phần dư của mô hình phân tích có trung bình = 0

# Hồi quy tuyến tính đơn biến

## Trình tự thực hiện:

- Nhập và lưu số liệu trong Excel.
- Chuyển số liệu vào phần mềm và tìm hiểu số liệu
- Xác định các thông số thống kê và kiểm định phân phối chuẩn.
- Vẽ đồ thị đám mây điểm và xem xét sự tương quan tuyến tính.
- Lập mô hình hồi quy tuyến tính và xem xét ý nghĩa thống kê của mô hình tuyến tính.
  - Kiểm tra các điều kiện cần thỏa của mô hình hồi quy tuyến tính.
  - Nếu tất cả các điều kiện đều thỏa; chấp nhận, giải thích kết quả mô hình hồi quy tuyến tính và vẽ đồ thị mô hình hồi quy tuyến tính.
  - Kết luận.

# Hồi quy tuyến tính đơn biến

## Kết luận

### Cách ghi kết quả Hồi quy tuyến tính đơn giản:

- Kết quả phân tích hồi tính tuyến tính đơn biến cho thấy biến age **giải thích được 87,75%** sự thay đổi nồng độ cholesterol trong máu của bệnh nhân [ $R^2 = 0,8775$ ,  $F(1, 16) = 114,6$ ,  $P\text{-value} = 1,06 \times 10^{-8} < \alpha = ?$ ]. Biến age **giải đoán được** nồng độ cholesterol trong máu [chặn intercept = 1,089;  $P\text{-value} = 0,00015 < \alpha = ?$ ; **độ dốc slope (age) = 0,058;  $P\text{-value} = 1,06 \times 10^{-8} < \alpha = ?$** ].

- Phương trình hồi quy dùng để nội suy **nồng độ cholesterol trong máu theo độ tuổi (chỉ áp dụng cho nam giới trong độ tuổi từ 20 (min) đến 63 (max))**:  $\text{choles} = 1,08922 + 0,05779 \times \text{age}$



# **GIỚI THIỆU**

## **THỰC TẬP ỨNG DỤNG THỐNG KÊ**

### **TRONG NGHIÊN CỨU SINH HỌC**

#### **CSH139 – SHH681 – BTE10019**

**Nguyễn Thị Gia Hằng**

**Đặng Lê Anh Tuấn**

**Nguyễn Trần Thụy Thanh Mai**

**TLTK: Thực tập Ứng dụng thống kê trong nghiên cứu Sinh học**

**Chuyên ngành: Sinh thái – Tài nguyên sinh vật**

**Bộ môn Sinh thái – Sinh học tiến hóa**

**Khoa Sinh học – Công nghệ sinh học**

**Trường Đại học Khoa học Tự nhiên – ĐHQG TpHCM**