Bài giảng 1: Giới thiệu về Thống kê sinh học

TS. Tô Đức Khánh

Khoa Toán-Tin Học, ĐH Khoa Học Tự Nhiên, Tp. HCM

Học phần: Thống Kê Trong Sinh Học

Nội dung buổi học

1 Một số thông tin của học phần

2 Mở đầu

3 Ôn lại một số kiến thức cơ bản

Thông tin của môn học

Giảng viên:

TS. Tô Đức Khánh (tôi)
Bộ môn Xác suất-Thống kê, khoa Toán-Tin học
Trường Đại Học Khoa Học Tự Nhiên, Tp. HCM
e-mail: tdkhanh@hcmus.edu.vn
Phòng F.012, Trường Đại Học Khoa Học Tự Nhiên, 227 Nguyễn Văn Cừ, Quận 5, Tp. Hồ Chí Minh.

Thông tin của môn học

Nội dung của môn học trong 14 tuần bao gồm:

- Bài toán thống kê sinh học, các cách lấy mẫu thông dụng;
- Các bài toán về ước lượng khoảng tin cậy;
- Các bài toán về kiểm định thống kê;
- Áp dụng mô hình hồi quy tuyến tính;
- Phân tích rủi ro và phân tích ROC;
- Giới thiệu mô hình hồi quy logistic và ứng dụng;
- Phân tích sống sót

Thực hành trên phần mềm R/Rstudio, với các thư viện: `tidyverse`, `ggplot2`, `glm`, `survival`, `pROC`.

Thông tin của môn học

Số tín chỉ

	Tín chỉ	Số giờ học
Lý thuyết	2	30
Thực hành R	1	30
Tổng cộng	3	60

Lịch học

- Tất cả các bài giảng lý thuyết, thực hành và bài tập được tổ chức tại cở sở 2 của trường Đại Học Khoa Học Tự Nhiên, Linh Trung, Thủ Đức.
- Bắt đầu từ 21/09/2023 tới 28/12/2023. Ngoại trừ ngày 16/11/2023 (thi giữa kỳ)
- Buổi lý thuyết: thứ 5 (hàng tuần) từ tiết 6 tới 8 (12:30 - 15:10) tại phòng D212, ngoại trừ các ngày 12/10/2023 tới 26/10/2023.
- Buổi thực hành R: thứ 6 (hàng tuần) từ tiết 2 tới 5 (09:30 - 11:30) tại phòng C204, bắt đầu từ tuần thứ 3 (05/10/2023).

Xem chi tiết lịch học tại trang quản lý môn học.

Thông tin của môn học

Cấu trúc điểm:

- Điểm quá trình thực hành: 20%
- Điểm thi giữa kỳ: 20%
- Điểm báo cáo nhóm: 10%
- Điểm thi cuối kỳ: 50%

Nhóm:
Sinh viên được chia thành các nhóm nhỏ, mỗi nhóm gồm từ 4-6 sinh viên.

Hình thức thi

- Thi giữa kỳ: viết tự luận trong thời gian 90 phút.
- Thi cuối kỳ: làm đề tài/bài tập lớn theo nhóm.

Chú ý: Sinh viên cần phải tham gia ít nhất 80% số tiết học.

Tài liệu học tập và tài liệu tham khảo

- Slide bài giảng được cập nhật tại trang của môn học: `https://courses.hcmus.edu.vn/course/view.php?id=3398`
- Motulsky, H. (2014). Intuitive biostatistics: a nonmathematical guide to statistical thinking. *Oxford University Press, USA.* **Tài liệu tham khảo**.
- Julien, J., & Hoffman, I. E. (2019). Basic Biostatistics for Medical and Biomedical Practitioners. *Academic Press*. **Tài liệu tham khảo**.
- Samuels, M. L., Witmer, J., & Schaffner, A. (2016). Statistics for the Life Sciences. *Pearson Education*. **Tài liệu tham khảo**.
- Hadley, W., & Garrett, G. R. (2017). R for data science: import, tidy, transform, visualize, and model data. *O'Reily Media Inc.,*. **Tài liệu tham khảo**.

1 Một số thông tin của học phần

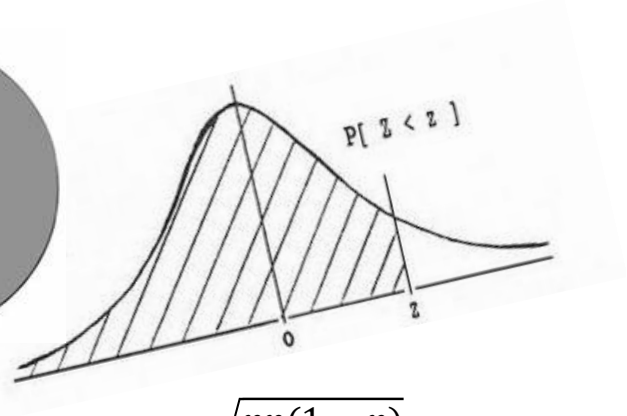
2 Mở đầu

3 Ôn lại một số kiến thức cơ bản

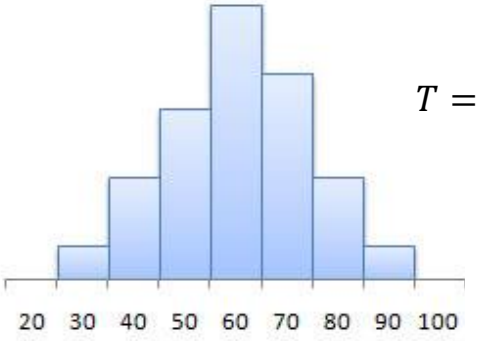
Thống kê trong sinh học?

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



p-value $= \Pr(|T| \geq t)$



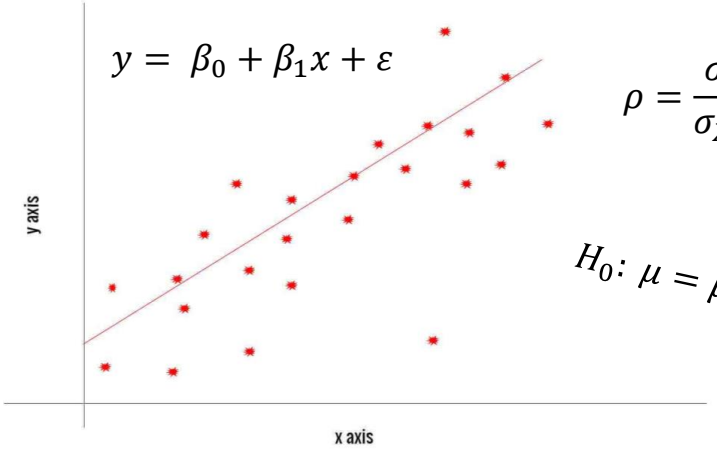
$$T = \frac{\bar{x} - \mu_0}{S_x/\sqrt{n}} \xrightarrow{d} t_{n-1}$$

$$\sqrt{np(1-p)}$$

$$\left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right)$$

$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$y = \beta_0 + \beta_1 x + \varepsilon$$



$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

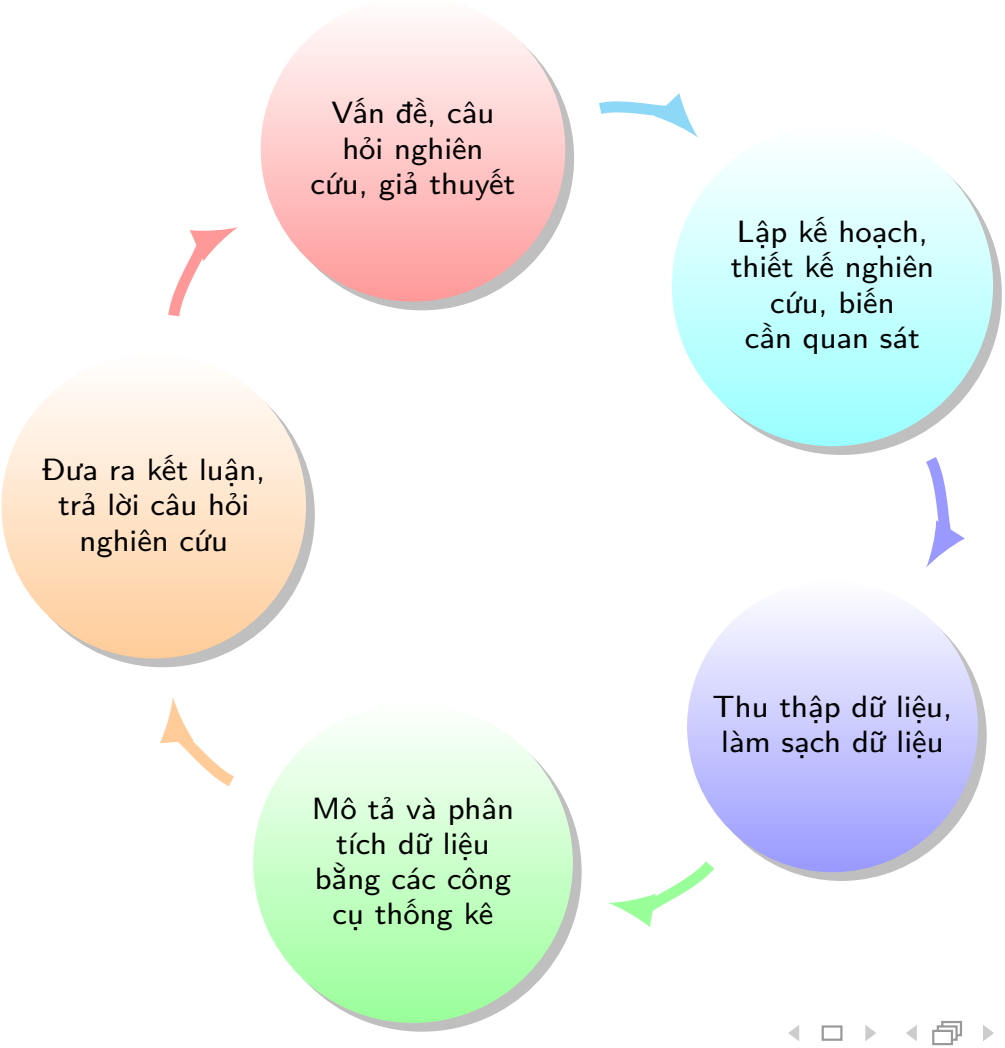
$$H_0: \mu = \mu_0$$

Thống kê trong sinh học?

Thống kê (**statistics**) là một ngành khoa học của việc thu thập, phân tích, biểu diễn và giải thích dữ liệu, nhằm tìm ra các bằng chứng để cung cố hoặc bác bỏ các giả thuyết khoa học.

Thống kê trong sinh học?

Thống kê (**statistics**) là một ngành khoa học của việc thu thập, phân tích, biểu diễn và giải thích dữ liệu, nhằm tìm ra các bằng chứng để cung cố hoặc bác bỏ các giả thuyết khoa học.



Thống kê trong sinh học?

Thống kê sinh học (**biostatistics**) là một nhánh của thống kê, mà ở đó nghiên cứu và vận dụng các phương pháp thống kê tới một loạt các chủ đề trong sinh học: sinh học phân tử, sinh học di truyền, y sinh, y tế, nhằm đưa ra các luận điểm/bằng chứng có **ý nghĩa thống kê** để củng cố hay bác bỏ các giả định khoa học chưa được kiểm định.

Câu hỏi:

- Câu hỏi nghiên cứu là gì?
- Lập kế hoạch cho một nghiên cứu thống kê là gì?
- Ý nghĩa thống kê là gì? và tại sao ta phải sử dụng các phương pháp thống kê trong các vấn đề của sinh học?

Câu hỏi nghiên cứu là gì?

Một câu hỏi nghiên cứu là "một câu hỏi mà một dự án nghiên cứu đặt ra để trả lời". Trong sinh học, câu hỏi nghiên cứu có thể là:

- sự khác biệt về hình thái giữa cua cái và cua đực, hay cua xanh và cam là gì?
- tại sao mức độ hoạt động của MAO (monoammino oxidase) lại khác nhau giữa các nhóm bệnh nhân tâm thần phân liệt?
- làm thế nào để phân loại các loài hoa iris dựa trên đặc điểm của chúng?
- có sự khác biệt nào giữa 3 loài chim cánh cụt: Adelie, Chinstrap và Gentoo?
- tỷ lệ 3:1 liệu có luôn luôn đúng trong di truyền?
- và nhiều hơn nữa ...

Lập kế hoạch nghiên cứu

Lập kế hoạch nghiên cứu là một bước rất quan trọng trong nghiên cứu phân tích dữ liệu. Ta cần làm rõ:

1. Đối tượng của nghiên cứu:
   - người;
   - động vật;
   - thực vật;
   - tế bào.
2. Khoanh vùng đối tượng:
   - khu vực sinh sống;
   - thu nhập;
   - cơ quan trên cơ thể (tim, não, cây, thân, lá);
   - tình trạng sức khỏe.
3. Các biến ngẫu nhiên cần được quan sát để phục vụ cho việc phân tích và trả lời câu hỏi nghiên cứu:
   - biến dạng số - định lượng (quantitative);
   - biến phân loại - định tính (qualitative).
4. Cách thức thu thập mẫu:
   - quan sát thực tế;
   - thí nghiệm.

Ví dụ 1: Monoamine oxidase (MAO)

Monoamine oxidase (MAO) v.s. bệnh tâm thần phân liệt (schizophrenia)

Monoamine oxidase (MAO) là một loại enzyme được cho là có vai trò trong việc điều chỉnh hành vi của não bộ, đặc biệt, nồng độ MAO có thể liên quan tới các tình trạng của bệnh tâm thần phân liệt (schizophrenia)[1].

- Câu hỏi nghiên cứu: MAO có liên quan tới tình trạng của bệnh nhân tâm thần phân liệt?
- Đối tượng nghiên cứu: bệnh nhân tâm thần phân liệt với các tình trạng khác nhau.
- Khoanh vùng đối tượng: Bệnh nhân tâm thần phân liệt tại một khu vực sinh sống nhất định.
- Biến cần quan sát:
  - nồng độ MAO của bệnh nhân tâm thần phân liệt;
  - tình trạng bệnh;
  - độ tuổi;
  - giới tính.
- Các thức lấy mẫu: quan sát thu thập dữ liệu thực tế.

[1] [N Engl J Med 1978; 298:61-66]

Ví dụ 2: Kích thước cơ thể và mức tiêu thụ năng lượng

Người to hơn cần nhiều năng lượng hơn?

Chúng ta thường nghĩ rằng người to hơn thì sẽ cần nhiều năng lượng hơn. Nhưng liệu điều này có đúng?

- Câu hỏi nghiên cứu: Nhu cầu đồ ăn có liên quan tới kích thước cơ thể.
- Đối tượng nghiên cứu: người trưởng thành.
- Khoanh vùng đối tượng: người trưởng thành tại một khu vực sinh sống nhất định.
- Biến cần quan sát:
  - trọng lượng cơ thể không có chất béo (the fat-free body mass);
  - năng lương tiêu thụ trong quá trình vận động tĩnh tại (suy nghĩ, đọc sách, nằm nghỉ);
  - tuổi;
  - giới tính.
- Các thức lấy mẫu: thí nghiệm và ghi nhận quan sát dữ liệu thực tế.

Ví dụ 3: thằn lằn sừng

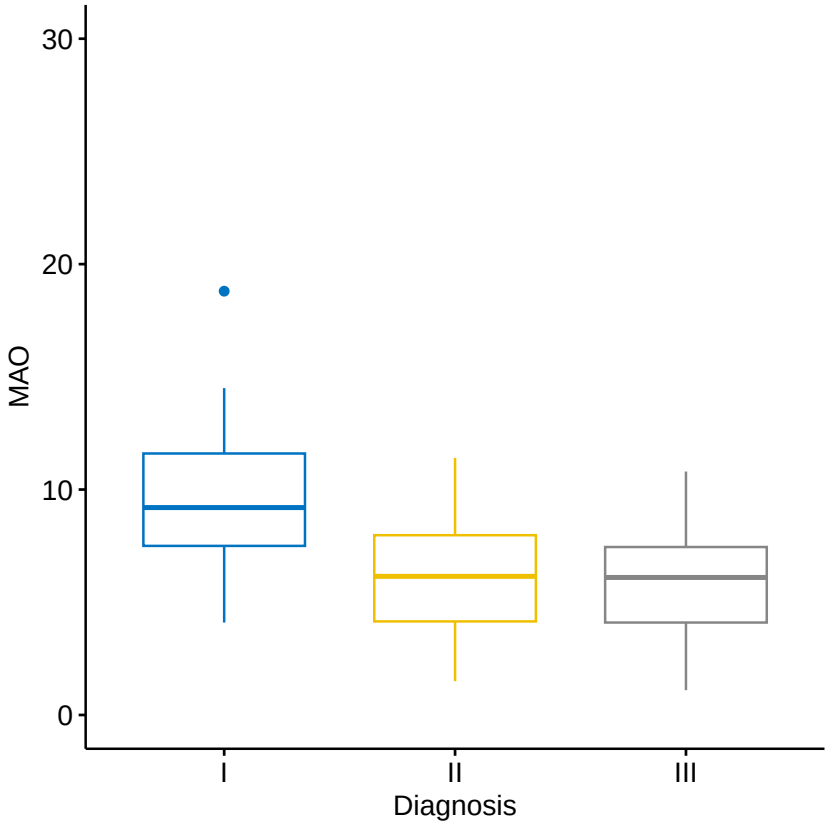
Chiều dài sừng của thằn lằn có giúp chúng thoát chết?

- Câu hỏi nghiên cứu: ???
- Đối tượng nghiên cứu: ???
- Khoanh vùng đối tượng: ???
- Biến cần quan sát: ???
- Các thức lấy mẫu: ???

Ý nghĩa thống kê?

So sánh sự khác biệt nồng độ của MAO trong tiểu cầu của 42 bệnh nhân mắc 3 loại bệnh tâm thần phân liệt:

- I - Tâm thần phân liệt không phân biệt mãn tính - chronic undifferentiated schizophrenia (18)
- II - Không phân biệt với các đặc điểm hoang tưởng - undifferentiated with paranoid features (16)
- III - Bệnh tâm thần phân liệt hoang tưởng - paranoid schizophrenia (8)



Ý nghĩa thống kê?

- Nồng độ MAO trong nhóm I, cao hơn rõ rệt (khoảng 50% số bệnh nhân trong nhóm I) so với nhóm II và III.
- Giữa nhóm II và III, sự khác biệt là không thực sự rõ ràng.
- Có vẻ như nồng độ thấp MAO có liên quan tới tình trạng bệnh II và III.

Ý nghĩa thống kê?

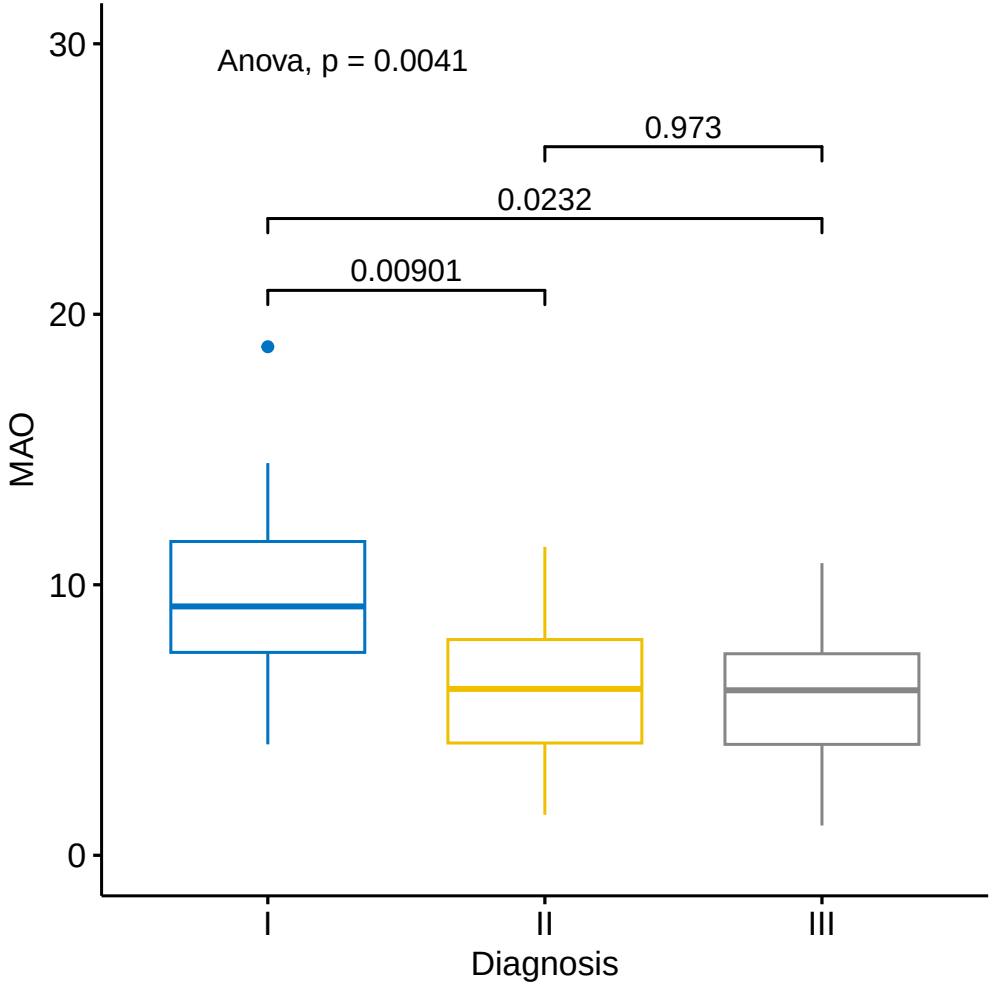
- Nồng độ MAO trong nhóm I, cao hơn rõ rệt (khoảng 50% số bệnh nhân trong nhóm I) so với nhóm II và III.
- Giữa nhóm II và III, sự khác biệt là không thực sự rõ ràng.
- Có vẻ như nồng độ thấp MAO có liên quan tới tình trạng bệnh II và III.

Vấn đề:

- Các kết luận trên có phải do **may mắn**???
- Mối liên hệ giữa MAO và tình trạng của bệnh tâm thần phân liệt có thực sự tồn tại thật???

Ý nghĩa thống kê?

Kết quả phân tích ANOVA và TukeyHSD (p-value).



Ý nghĩa thống kê?

Một cách *bản năng*, so sánh *p*-value với một mức ý nghĩa (*significance level*), chẳng hạn 0.05, ta dễ dàng đưa ra một số nhận định thống kê như sau:

- Một cách tổng thể, sự khác biệt nồng độ MAO trong ba nhóm bệnh là có **ý nghĩa thống kê**.
- Sự khác biệt về nồng độ MAO giữa nhóm I và II, và giữa nhóm I và III là có **ý nghĩa thống kê**.
- Không tìm thấy **ý nghĩa thống kê** trong so sánh nồng độ MAO giữa nhóm II và III.

Ý nghĩa thống kê?

Một cách *bản năng*, so sánh *p*-value với một mức ý nghĩa (*significance level*), chẳng hạn 0.05, ta dễ dàng đưa ra một số nhận định thống kê như sau:

- Một cách tổng thể, sự khác biệt nồng độ MAO trong ba nhóm bệnh là có **ý nghĩa thống kê**.
- Sự khác biệt về nồng độ MAO giữa nhóm I và II, và giữa nhóm I và III là có **ý nghĩa thống kê**.
- Không tìm thấy **ý nghĩa thống kê** trong so sánh nồng độ MAO giữa nhóm II và III.

Như vậy:

- **Ý nghĩa thống kê** là sự quyết định của nhà phân tích thống kê rằng các kết quả trong dữ liệu không thể giải thích được một cách ngẫu nhiên.
- **Kiểm tra giả thuyết thống kê** là phương pháp mà nhà phân tích đưa ra quyết định này.

1 Một số thông tin của học phần

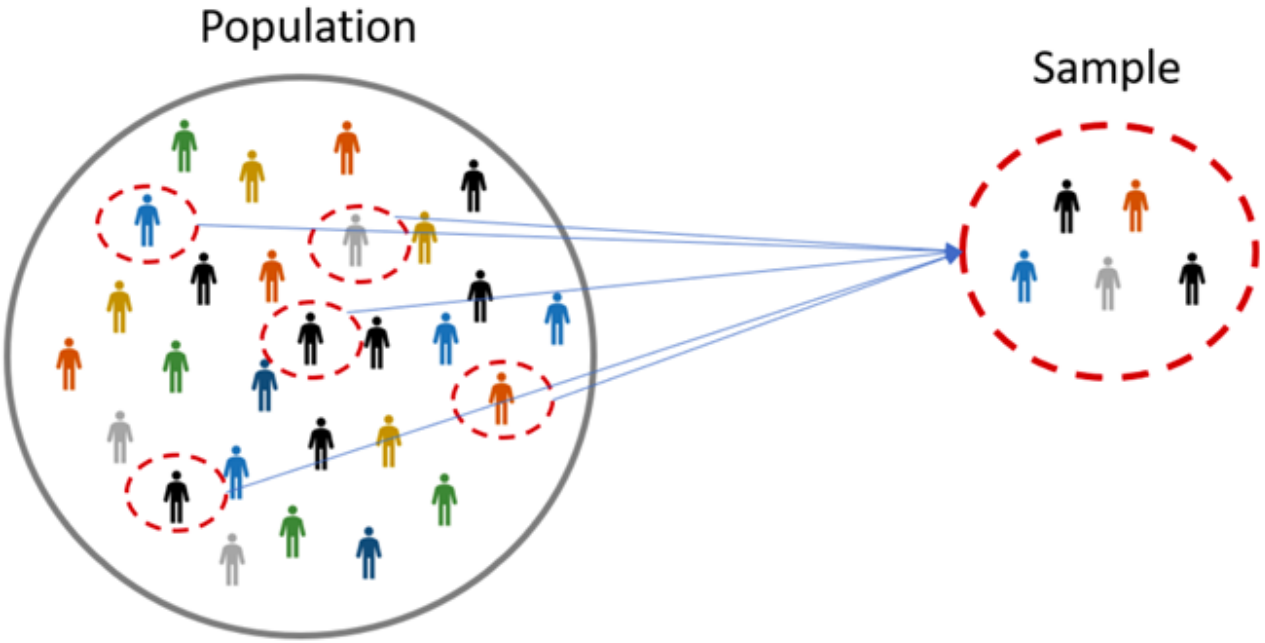
2 Mở đầu

3 Ôn lại một số kiến thức cơ bản

Dữ liệu thống kê là gì?

- Một dữ liệu thống kê là một bộ *mẫu ngẫu nhiên* của các *đối tượng nghiên cứu*, nó bao gồm các thông tin mô tả *tính chất của các đối tượng nghiên cứu*.
- Một bộ dữ liệu thống kê có thể được thu được thông qua việc quan sát và thu mẫu thực tế (*observations*), hoặc thu mẫu các đối tượng trong phòng thí nghiệm *experiments*.
- Một quần thể (*population*) là bộ dữ liệu lớn bao gồm tất cả đối tượng nghiên cứu có thể được bao hàm trong câu hỏi nghiên cứu.
- Chú ý: một đối tượng nghiên cứu được coi là một đơn vị thống kê (*statistical unit*).

Dữ liệu thống kê là gì?



Dữ liệu thống kê là gì?

Characteristics of a iris flower

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

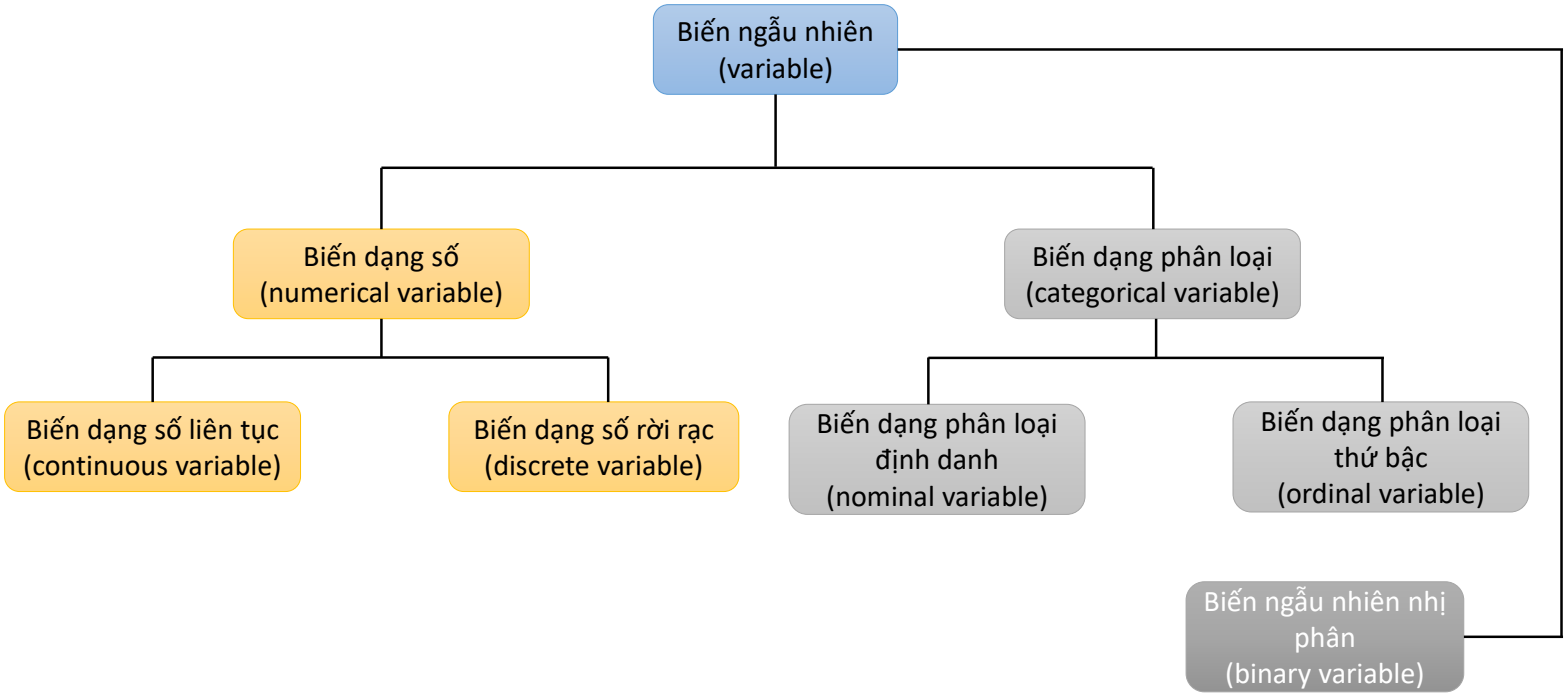
Number of iris flowers

- Câu hỏi nghiên cứu: Làm thế nào phân loại 3 loài hoa diên vĩ (iris): setosa, versicolor và virginica, dựa vào các số đo của chiều dài đài hoa (sepal), chiều rông đài hoa, chiều dài cánh hoa (petal) và chiều rộng cánh hoa?
- Đối tượng nghiên cứu: hoa diên vĩ thuộc 3 loài: setosa, versicolor, và virginica.
- Quần thể: Tất cả các hoa diên vĩ trên thế giới, với khoảng 310 loài đã được ghi nhận.
- Dữ liệu: một mẫu gồm 150 hoa diên vĩ của 3 loài: setosa, versicolor, and virginica.

Biến ngẫu nhiên là gì?

Biến ngẫu nhiên

Trong thống kê, một tính chất của đối tượng nghiên cứu thường được gán bởi một biến ngẫu nhiên (*random variable*).



Biến ngẫu nhiên là gì?

Biến ngẫu nhiên dạng số (*numerical variable* hay *quantitative variable*) là biến ngẫu nhiên với giá trị là dạng:

- liên tục (*continuous variable*);
- rời rạc hay số đếm (*discrete variable*),

ghi lại số lượng của một đối đối tượng nào đó.

Biến ngẫu nhiên dạng phân loại (*categorical variable* hay *qualitative variable*) là biến ngẫu nhiên có giá trị là các thể loại khác nhau của người hoặc vật. Thông thường ta có hai dạng phổ biến sau:

- biến định danh (*nominal variable*)
- biến thứ bậc (*ordinal variable*)

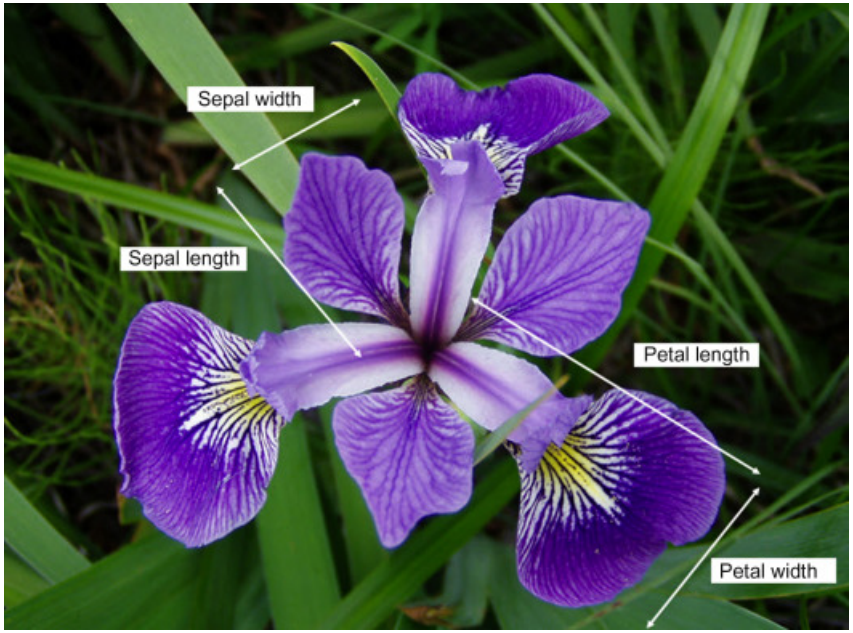
Đặc biệt, đối với biến định danh, ta có:

- biến nhị định danh hay biến nhị phân (*binary variable*), khi biến có 2 giá trị;
- biến đa định danh hay biến đa thức (*nominal multinomial variable* hoặc *multinomial variable*).

Biến ngẫu nhiên là gì?

Ví dụ, đối với dữ liệu hoa diên vĩ, ta có:

- 04 biến dạng số liên tục: chiều dài đài hoa, chiều rộng đài hoa, chiều dài cánh hoa, chiều rộng cánh hoa;
- 01 biến dạng phân loại định danh: loài hoa.



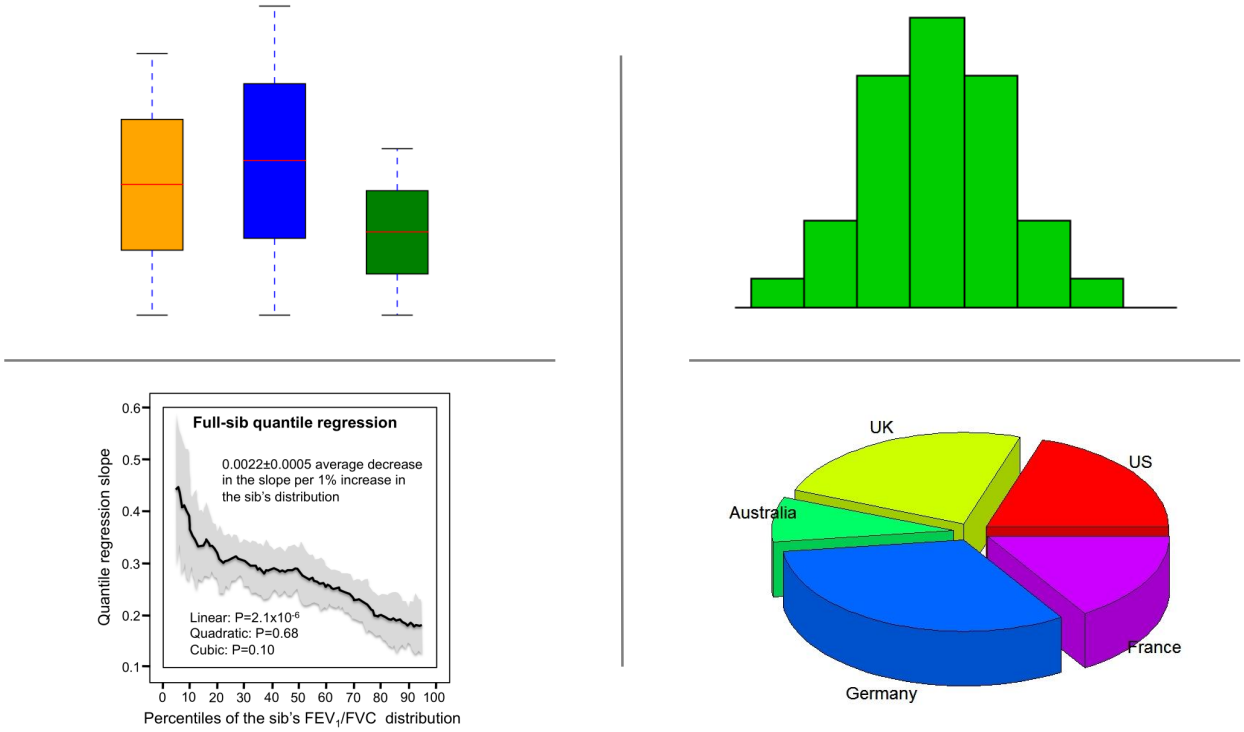
Biến ngẫu nhiên là gì?

Câu hỏi: phân loại dạng biến ngẫu nhiên

- Các cách thức truyền nhiễm bệnh.
- Giới tính sinh học của một người.
- Chiều cao của một người.
- Độ tuổi chết của một người.
- Các nhóm tính trạng bệnh của bệnh ung thư phổi.
- Lượng thuốc lá tiêu thụ trong một ngày.

Kỹ thuật phân tích thống kê

Thống kê mô tả (*descriptive statistics*) là tổ hợp các kỹ thuật để tổ chức, mô tả, tổng hợp và trình bày dữ liệu hoặc thông tin của dữ liệu theo cách thuận tiện cho việc hiểu và sử dụng.



Kỹ thuật phân tích thống kê

Thống kê suy luận (*inferential statistics*) là tổ hợp các phương pháp thống kê dùng để đưa ra các kết luận và suy luận về tính chất của quẩn thể đối tượng dựa trên thông tin của dữ liệu.

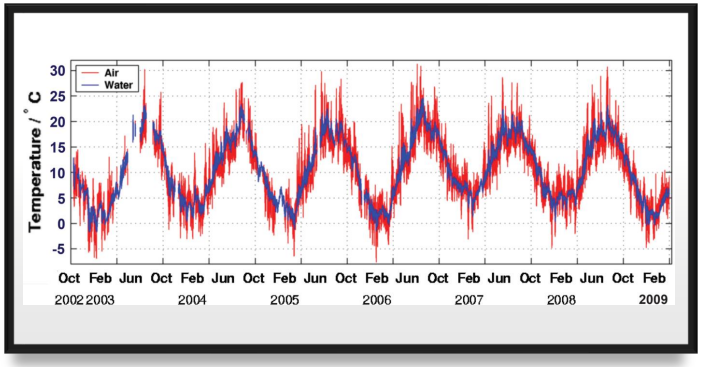
- Khoảng tin cậy cho trung bình quần thể.
- Kiểm định giả thuyết.
- So sánh nhiều nhóm trong một quần thể.
- Dự đoán xu hướng.

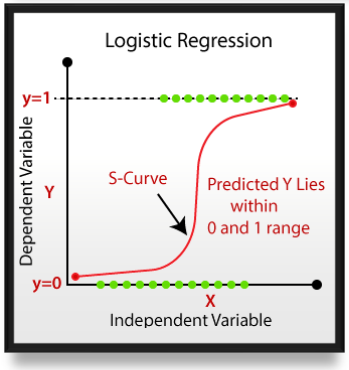
Kỹ thuật phân tích thống kê

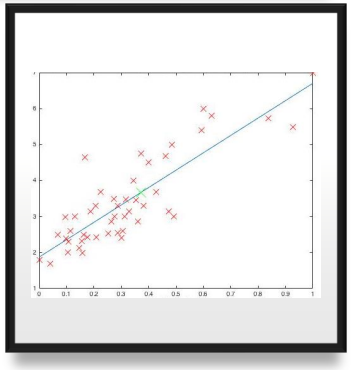
Mô hình thống kê (*statistical model*) là một dạng phân tích thống kê dựa trên một mô hình toán cụ thể cho dữ liệu. Mô hình thống kê thường được dùng để mô tả sự biến động của một (hoặc nhiều) biến dựa trên giá trị của một hay nhiều biến khác. Khi phân tích mô hình thống kê, ta thường phải sửa dụng một số phương pháp thống kê mô tả và thống kê suy luận.

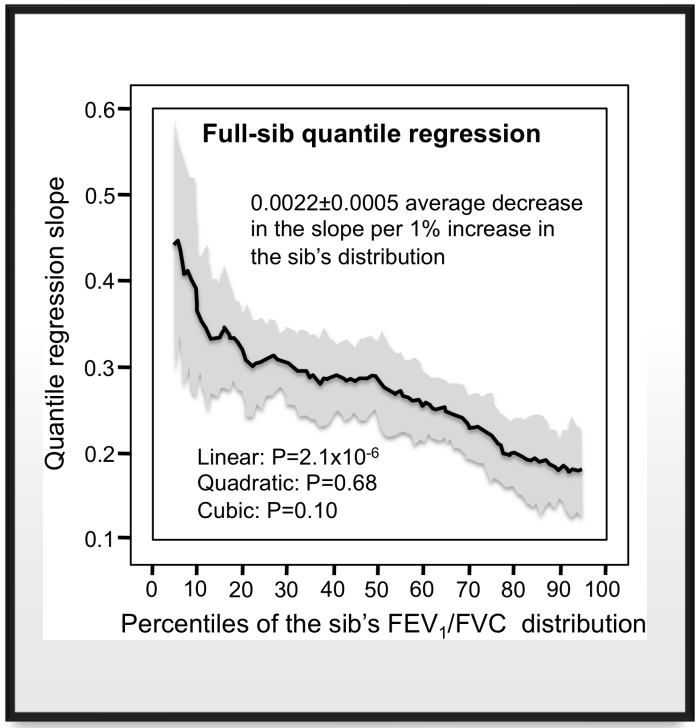
- Mô hình hồi quy tuyến tính.
- Mô hình hồi quy logistic.
- Mô hình hồi quy phân vị (quantile regression model).
- Mô hình chuỗi thời gian.
- Mô hình sống sót.

Kỹ thuật phân tích thống kê



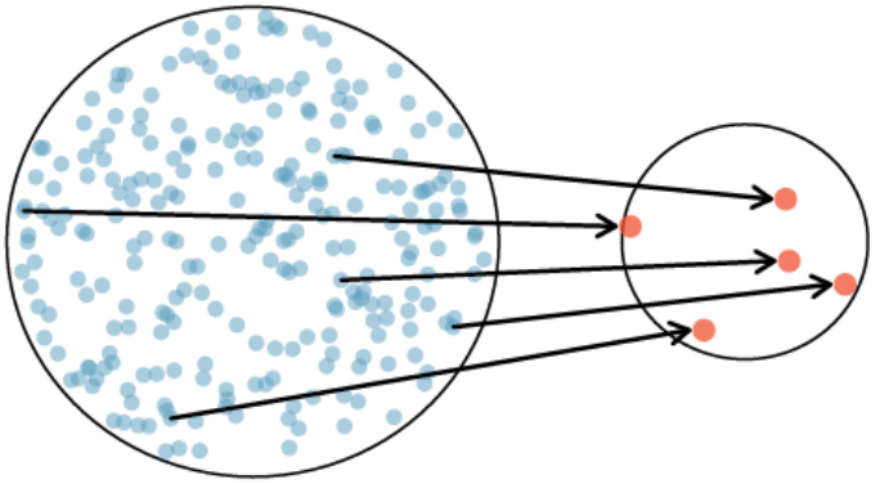






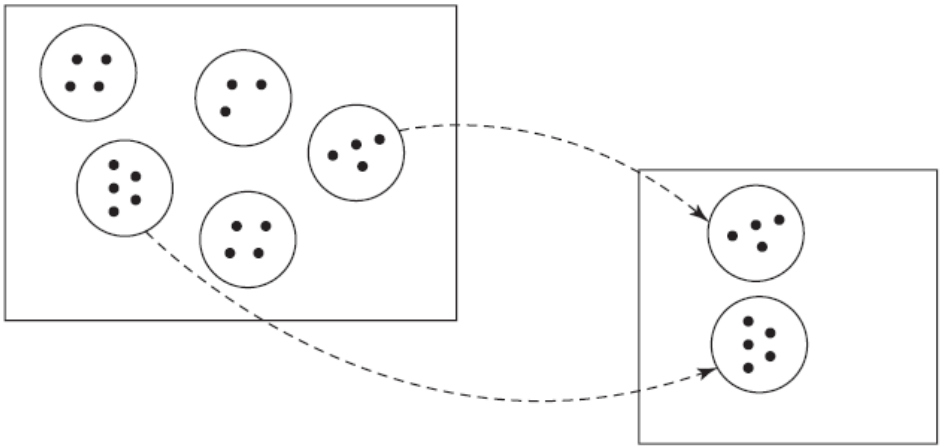
Kỹ thuật lấy mẫu ngẫu nhiên

Lấy mẫu ngẫu nhiên đơn giản (*simple random sampling*) là một cách lấy mẫu ngẫu nhiên mà trong đó, ta chọn ngẫu nhiên n đối tượng từ quần thể sao cho mỗi đối tượng một lần với xác suất được chọn là như nhau, các đối tượng là độc lập nhau.



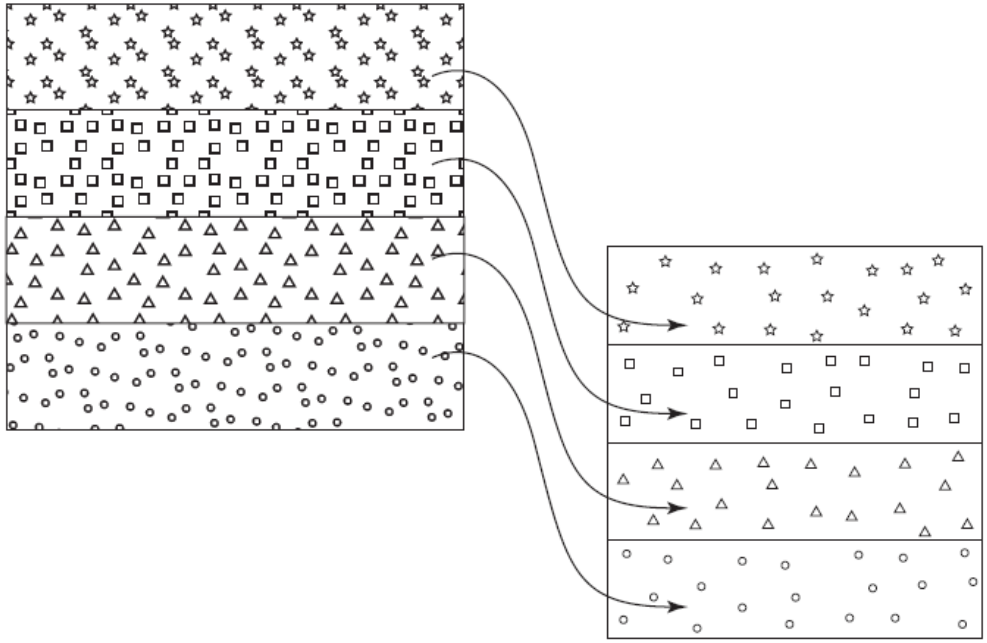
Kỹ thuật lấy mẫu ngẫu nhiên

Lấy mẫu ngẫu nhiên theo cụm (*random cluster sample*) là một cách lấy mẫu ngẫu nhiên mà trong đó, ta sẽ chọn ngẫu nhiên **toàn bộ** thành viên trong một nhóm thay vì chỉ định một thành viên duy nhất.



Kỹ thuật lấy mẫu ngẫu nhiên

Lấy mẫu ngẫu nhiên phân tầng (*stratified random sample*) một mẫu ngẫu nhiên phân tầng được chọn bằng cách trước tiên chia quần thể thành các tầng, tức là tập hợp các cá thể đồng nhất. Sau đó, nhiều mẫu ngẫu nhiên đơn giản được lấy - một mẫu trong mỗi tầng - và kết hợp lại để tạo thành mẫu.



Các sai số khi lấy mẫu

Khi thực hiện lấy mẫu, ta có thể gặp phải một trong số các lỗi/sai số lấy mẫu sau:

Sai số mẫu (*sampling error*) hay *sai số lấy mẫu ngẫu nhiên* xảy ra khi mẫu được chọn bằng phương pháp lấy mẫu ngẫu nhiên nhưng có sự khác biệt giữa kết quả mẫu và kết quả tổng thể thực; sai số này là kết quả của sự biến động ngẫu nhiên của mẫu.

Sai số không lấy mẫu (*nonsampling error*) là kết quả của lỗi của con người, bao gồm các yếu tố như nhập dữ liệu sai, lỗi tính toán, câu hỏi có từ ngữ sai lệch, dữ liệu sai do người trả lời cung cấp, đưa ra kết luận sai lệch hoặc áp dụng các phương pháp thống kê không phù hợp với hoàn cảnh.

Sai số lấy mẫu không ngẫu nhiên là kết quả của việc sử dụng phương pháp lấy mẫu không ngẫu nhiên, chẳng hạn như sử dụng mẫu thuận tiện.

Bài tập

Trong mỗi nghiên cứu sau đây, hãy xác định kỹ thuật lấy mẫu nào mô tả đúng nhất cách thu thập dữ liệu (hoặc có thể được xử lý như thể chúng được thu thập): lấy mẫu ngẫu nhiên đơn giản, lấy mẫu cụm ngẫu nhiên hoặc lấy mẫu ngẫu nhiên phân tầng. Đối với lấy mẫu theo cụm, hãy xác định các cụm và đối với lấy mẫu phân tầng, hãy xác định các tầng.

a. Tất cả 257 bệnh nhân mắc bệnh bạch cầu từ ba phòng khám nhi được chọn ngẫu nhiên ở Hoa Kỳ đã được ghi danh vào thử nghiệm lâm sàng một loại thuốc mới.

b. Tổng cộng có 12 mẫu đất nặng 10 g được thu thập từ các vị trí ngẫu nhiên trong một trang trại để nghiên cứu các đặc tính vật lý và hóa học của đất.

c. Trong một nghiên cứu về ô nhiễm, ba mẫu không khí 100 ml được thu thập ở mỗi độ cao trong số bốn độ cao cụ thể (100 m, 500 m, 1000 m, 2000 m) để có tổng số 12 mẫu 100 ml.

d. Tổng cộng có 20 quả nho riêng lẻ được hái, mỗi quả từ 20 cây nho ngẫu nhiên trong vườn nho, để đánh giá mức độ sẵn sàng cho thu hoạch.

e. 24 con chó (tám giống chó nhỏ được chọn ngẫu nhiên, tám giống chó trung bình được chọn ngẫu nhiên và tám giống chó lớn được chọn ngẫu nhiên) đã được ghi danh vào một thí nghiệm để đánh giá một chương trình huấn luyện mới.

Một số chủ đề thảo luận cho tuần sau:

- p-value là gì? sử dụng như thế nào cho hợp lý.
- Một số ví dụ cho các cách lấy mẫu ngẫu nhiên.