TRƯỜNG ĐH CÔNG NGHỆ THÔNG TIN

BÀI GIẢNG
XÁC SUẤT THỐNG KÊ

Giảng viên: TS. PHÙNG MINH ĐỨC

(Bộ môn Toán Lý)

Chương 5: Thống kê toán học

5.1 Lý thuyết mẫu

Ví dụ

Ta muốn biết thu nhập trung bình trong năm 2021 của giáo viên đang giảng dạy ở Thành phố Hồ Chí Minh.

Ví dụ

Ta muốn biết thu nhập trung bình trong năm 2021 của giáo viên đang giảng dạy ở Thành phố Hồ Chí Minh. Ta sẽ lập danh sách tất cả các giáo viên đang dạy ở thành phố Hồ Chí Minh và ghi lại thu nhập của từng người trong năm 2021.

Ví dụ

Ta muốn biết thu nhập trung bình trong năm 2021 của giáo viên đang giảng dạy ở Thành phố Hồ Chí Minh. Ta sẽ lập danh sách tất cả các giáo viên đang dạy ở thành phố Hồ Chí Minh và ghi lại thu nhập của từng người trong năm 2021.
Tuy nhiên, việc thu thập dữ liệu của tất cả các giáo viên tốn rất nhiều thời gian, công sức.

Ví dụ

Ta muốn biết thu nhập trung bình trong năm 2021 của giáo viên đang giảng dạy ở Thành phố Hồ Chí Minh. Ta sẽ lập danh sách tất cả các giáo viên đang dạy ở thành phố Hồ Chí Minh và ghi lại thu nhập của từng người trong năm 2021.
Tuy nhiên, việc thu thập dữ liệu của tất cả các giáo viên tốn rất nhiều thời gian, công sức. Do đó, người ta có thể chọn ra một nhóm giáo viên (ta sẽ gọi là **mẫu**) trong toàn bộ giáo viên (gọi là **tổng thể**) để điều tra.

Ví dụ

Ta muốn biết thu nhập trung bình trong năm 2021 của giáo viên đang giảng dạy ở Thành phố Hồ Chí Minh. Ta sẽ lập danh sách tất cả các giáo viên đang dạy ở thành phố Hồ Chí Minh và ghi lại thu nhập của từng người trong năm 2021.
Tuy nhiên, việc thu thập dữ liệu của tất cả các giáo viên tốn rất nhiều thời gian, công sức. Do đó, người ta có thể chọn ra một nhóm giáo viên (ta sẽ gọi là **mẫu**) trong toàn bộ giáo viên (gọi là **tổng thể**) để điều tra. Bằng các phương pháp của ngành thống kê, người ta có thể đưa ra được mức thu nhập trung bình của toàn bộ giáo viên trên địa bàn Thành phố Hồ Chí Minh.

Ví dụ

Ta muốn biết thu nhập trung bình trong năm 2021 của giáo viên đang giảng dạy ở Thành phố Hồ Chí Minh. Ta sẽ lập danh sách tất cả các giáo viên đang dạy ở thành phố Hồ Chí Minh và ghi lại thu nhập của từng người trong năm 2021.
Tuy nhiên, việc thu thập dữ liệu của tất cả các giáo viên tốn rất nhiều thời gian, công sức. Do đó, người ta có thể chọn ra một nhóm giáo viên (ta sẽ gọi là **mẫu**) trong toàn bộ giáo viên (gọi là **tổng thể**) để điều tra. Bằng các phương pháp của ngành thống kê, người ta có thể đưa ra được mức thu nhập trung bình của toàn bộ giáo viên trên địa bàn Thành phố Hồ Chí Minh.

Chúng ta quan tâm đến tổng thể, nhưng tổng thể có thể khó hoặc không thể thống kê được.

Ví dụ

Ta muốn biết thu nhập trung bình trong năm 2021 của giáo viên đang giảng dạy ở Thành phố Hồ Chí Minh. Ta sẽ lập danh sách tất cả các giáo viên đang dạy ở thành phố Hồ Chí Minh và ghi lại thu nhập của từng người trong năm 2021.
Tuy nhiên, việc thu thập dữ liệu của tất cả các giáo viên tốn rất nhiều thời gian, công sức. Do đó, người ta có thể chọn ra một nhóm giáo viên (ta sẽ gọi là **mẫu**) trong toàn bộ giáo viên (gọi là **tổng thể**) để điều tra. Bằng các phương pháp của ngành thống kê, người ta có thể đưa ra được mức thu nhập trung bình của toàn bộ giáo viên trên địa bàn Thành phố Hồ Chí Minh.

Chúng ta quan tâm đến tổng thể, nhưng tổng thể có thể khó hoặc không thể thống kê được. Thay vào đó, **người ta cố gắng mô tả hoặc dự đoán các đặc điểm của tổng thể trên cơ sở thông tin thu được từ mẫu đại diện của tổng thể đó.**

5.1.1 Giới thiệu về mẫu

5.1.1 Giới thiệu về mẫu

Định nghĩa 5.1.1

1. Một **tổng thể** (population) là tập hợp tất cả các đối tượng có chung một tính chất mà ta quan tâm. Số phần tử của tổng thể được gọi là *kích thước tổng thể*.

5.1.1 Giới thiệu về mẫu

Định nghĩa 5.1.1

1. Một **tổng thể** (population) là tập hợp tất cả các đối tượng có chung một tính chất mà ta quan tâm. Số phần tử của tổng thể được gọi là *kích thước tổng thể*.
2. Việc chọn từ tổng thể một tập con nào đó được gọi là **phép lấy mẫu**.

5.1.1 Giới thiệu về mẫu

Định nghĩa 5.1.1

1. Một **tổng thể** (population) là tập hợp tất cả các đối tượng có chung một tính chất mà ta quan tâm. Số phần tử của tổng thể được gọi là *kích thước tổng thể*.
2. Việc chọn từ tổng thể một tập con nào đó được gọi là **phép lấy mẫu**.
3. Một **mẫu** (sample) là một tập con của tổng thể. Số phần tử của mẫu được gọi là *kích thước mẫu*.

5.1.1 Giới thiệu về mẫu

Định nghĩa 5.1.1

1. Một **tổng thể** (population) là tập hợp tất cả các đối tượng có chung một tính chất mà ta quan tâm. Số phần tử của tổng thể được gọi là *kích thước tổng thể*.
2. Việc chọn từ tổng thể một tập con nào đó được gọi là **phép lấy mẫu**.
3. Một **mẫu** (sample) là một tập con của tổng thể. Số phần tử của mẫu được gọi là *kích thước mẫu*.

Ví dụ 5.1.2 Một công ty sản xuất chip máy tính đóng gói mỗi hộp gồm 100 chip. Người ta muốn khảo sát tỉ lệ chip bị lỗi trong một lô hàng gồm 1000 hộp của công ty. Chọn ngẫu nhiên 80 hộp chip để kiểm tra.

5.1.1 Giới thiệu về mẫu

Định nghĩa 5.1.1

1. Một **tổng thể** (population) là tập hợp tất cả các đối tượng có chung một tính chất mà ta quan tâm. Số phần tử của tổng thể được gọi là *kích thước tổng thể*.
2. Việc chọn từ tổng thể một tập con nào đó được gọi là **phép lấy mẫu**.
3. Một **mẫu** (sample) là một tập con của tổng thể. Số phần tử của mẫu được gọi là *kích thước mẫu*.

Ví dụ 5.1.2 Một công ty sản xuất chip máy tính đóng gói mỗi hộp gồm 100 chip. Người ta muốn khảo sát tỉ lệ chip bị lỗi trong một lô hàng gồm 1000 hộp của công ty. Chọn ngẫu nhiên 80 hộp chip để kiểm tra. $\Rightarrow$ Ta có **tổng thể** là 1000 hộp chip và **mẫu** là 80 hộp chip được kiểm tra.

Một trong những nhiệm vụ quan trọng của thống kê là xây dựng các phương pháp cho phép rút ra các kết luận hoặc đưa ra dự báo về toàn bộ tổng thể dựa trên một mẫu. Do đó, vấn đề lấy mẫu là một việc vô cùng quan trọng.

Một trong những nhiệm vụ quan trọng của thống kê là xây dựng các phương pháp cho phép rút ra các kết luận hoặc đưa ra dự báo về toàn bộ tổng thể dựa trên một mẫu. Do đó, vấn đề lấy mẫu là một việc vô cùng quan trọng.

Định nghĩa 5.1.3 (Mẫu ngẫu nhiên)

Một mẫu là ngẫu nhiên (random sample) nếu trong phép lấy mẫu đó, mỗi phần tử được chọn một cách độc lập và có xác suất được chọn như nhau.

Một trong những nhiệm vụ quan trọng của thống kê là xây dựng các phương pháp cho phép rút ra các kết luận hoặc đưa ra dự báo về toàn bộ tổng thể dựa trên một mẫu. Do đó, vấn đề lấy mẫu là một việc vô cùng quan trọng.

Định nghĩa 5.1.3 (Mẫu ngẫu nhiên)

Một mẫu là ngẫu nhiên (random sample) nếu trong phép lấy mẫu đó, mỗi phần tử được chọn một cách độc lập và có xác suất được chọn như nhau.

Yêu cầu khi lấy mẫu:

Một trong những nhiệm vụ quan trọng của thống kê là xây dựng các phương pháp cho phép rút ra các kết luận hoặc đưa ra dự báo về toàn bộ tổng thể dựa trên một mẫu. Do đó, vấn đề lấy mẫu là một việc vô cùng quan trọng.

Định nghĩa 5.1.3 (Mẫu ngẫu nhiên)

Một mẫu là ngẫu nhiên (random sample) nếu trong phép lấy mẫu đó, mỗi phần tử được chọn một cách độc lập và có xác suất được chọn như nhau.

Yêu cầu khi lấy mẫu:

- Mẫu được chọn là mẫu ngẫu nhiên.

Một trong những nhiệm vụ quan trọng của thống kê là xây dựng các phương pháp cho phép rút ra các kết luận hoặc đưa ra dự báo về toàn bộ tổng thể dựa trên một mẫu. Do đó, vấn đề lấy mẫu là một việc vô cùng quan trọng.

Định nghĩa 5.1.3 (Mẫu ngẫu nhiên)

Một mẫu là ngẫu nhiên (random sample) nếu trong phép lấy mẫu đó, mỗi phần tử được chọn một cách độc lập và có xác suất được chọn như nhau.

Yêu cầu khi lấy mẫu:

- Mẫu được chọn là mẫu ngẫu nhiên.
- Kích thước mẫu đủ lớn.

Một trong những nhiệm vụ quan trọng của thống kê là xây dựng các phương pháp cho phép rút ra các kết luận hoặc đưa ra dự báo về toàn bộ tổng thể dựa trên một mẫu. Do đó, vấn đề lấy mẫu là một việc vô cùng quan trọng.

Định nghĩa 5.1.3 (Mẫu ngẫu nhiên)

Một mẫu là ngẫu nhiên (random sample) nếu trong phép lấy mẫu đó, mỗi phần tử được chọn một cách độc lập và có xác suất được chọn như nhau.

Yêu cầu khi lấy mẫu:

- Mẫu được chọn là mẫu ngẫu nhiên.
- Kích thước mẫu đủ lớn.

Nếu kích thước mẫu càng lớn thì thông tin suy luận về tổng thể càng đáng tin cậy và có ý nghĩa.

Ví dụ

Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố.

Ví dụ

Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố.

- Có 2 nhóm khảo sát:
  1. Nhóm 1 khảo sát một mẫu ngẫu nhiên gồm 50 người và tìm được chiều cao trung bình của 50 người này là 164 cm.

Ví dụ

Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố.

- Có 2 nhóm khảo sát:
  1. Nhóm 1 khảo sát một mẫu ngẫu nhiên gồm 50 người và tìm được chiều cao trung bình của 50 người này là 164 cm.
  2. Nhóm 2 khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm.

Ví dụ

Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố.

- Có 2 nhóm khảo sát:
  1. Nhóm 1 khảo sát một mẫu ngẫu nhiên gồm 50 người và tìm được chiều cao trung bình của 50 người này là 164 cm.
  2. Nhóm 2 khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm.
     Khi đó suy luận về chiều cao trung bình của tổng thể của nhóm 2 đáng tin cậy hơn nhóm 1.

Ví dụ

Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố.

- Có 2 nhóm khảo sát:
  1. Nhóm 1 khảo sát một mẫu ngẫu nhiên gồm 50 người và tìm được chiều cao trung bình của 50 người này là 164 cm.
  2. Nhóm 2 khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm.
     Khi đó suy luận về chiều cao trung bình của tổng thể của nhóm 2 đáng tin cậy hơn nhóm 1.
- Một nhóm khảo sát khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm.

Ví dụ

Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố.

- Có 2 nhóm khảo sát:
  1. Nhóm 1 khảo sát một mẫu ngẫu nhiên gồm 50 người và tìm được chiều cao trung bình của 50 người này là 164 cm.
  2. Nhóm 2 khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm.
     Khi đó suy luận về chiều cao trung bình của tổng thể của nhóm 2 đáng tin cậy hơn nhóm 1.
- Một nhóm khảo sát khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm. Khi đó nhóm đưa ra các kết luận
  1. Chiều cao trung bình của tổng thể là 164,6 cm.

Ví dụ

Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố.

- Có 2 nhóm khảo sát:
  1. Nhóm 1 khảo sát một mẫu ngẫu nhiên gồm 50 người và tìm được chiều cao trung bình của 50 người này là 164 cm.
  2. Nhóm 2 khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm.
     Khi đó suy luận về chiều cao trung bình của tổng thể của nhóm 2 đáng tin cậy hơn nhóm 1.
- Một nhóm khảo sát khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm. Khi đó nhóm đưa ra các kết luận
  1. Chiều cao trung bình của tổng thể là 164,6 cm.
  2. Chiều cao trung bình của tổng thể là từ 164 cm đến 165 cm.

Ví dụ

Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố.

- Có 2 nhóm khảo sát:
  1. Nhóm 1 khảo sát một mẫu ngẫu nhiên gồm 50 người và tìm được chiều cao trung bình của 50 người này là 164 cm.
  2. Nhóm 2 khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm.
     Khi đó suy luận về chiều cao trung bình của tổng thể của nhóm 2 đáng tin cậy hơn nhóm 1.
- Một nhóm khảo sát khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm. Khi đó nhóm đưa ra các kết luận
  1. Chiều cao trung bình của tổng thể là 164,6 cm.
  2. Chiều cao trung bình của tổng thể là từ 164 cm đến 165 cm.
  3. Chiều cao trung bình của tổng thể là từ 163,5 cm đến 165,5 cm.

Ví dụ

Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố.

- Có 2 nhóm khảo sát:
  1. Nhóm 1 khảo sát một mẫu ngẫu nhiên gồm 50 người và tìm được chiều cao trung bình của 50 người này là 164 cm.
  2. Nhóm 2 khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm.
     Khi đó suy luận về chiều cao trung bình của tổng thể của nhóm 2 đáng tin cậy hơn nhóm 1.
- Một nhóm khảo sát khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm. Khi đó nhóm đưa ra các kết luận
  1. Chiều cao trung bình của tổng thể là 164,6 cm.
  2. Chiều cao trung bình của tổng thể là từ 164 cm đến 165 cm.
  3. Chiều cao trung bình của tổng thể là từ 163,5 cm đến 165,5 cm.
  4. Chiều cao trung bình của tổng thể là từ 130 cm đến 185 cm.

Ví dụ

Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố.

- Có 2 nhóm khảo sát:
  1. Nhóm 1 khảo sát một mẫu ngẫu nhiên gồm 50 người và tìm được chiều cao trung bình của 50 người này là 164 cm.
  2. Nhóm 2 khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm.
     Khi đó suy luận về chiều cao trung bình của tổng thể của nhóm 2 đáng tin cậy hơn nhóm 1.
- Một nhóm khảo sát khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm. Khi đó nhóm đưa ra các kết luận
  1. Chiều cao trung bình của tổng thể là 164,6 cm.
  2. Chiều cao trung bình của tổng thể là từ 164 cm đến 165 cm.
  3. Chiều cao trung bình của tổng thể là từ 163,5 cm đến 165,5 cm.
  4. Chiều cao trung bình của tổng thể là từ 130 cm đến 185 cm.

  Trong 3 suy luận đầu tiên, suy luận nào **đáng tin cậy hơn**? (tức là có thể lấy làm kết quả chung cho toàn thành phố)

Ví dụ

Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố.

- Có 2 nhóm khảo sát:
  1. Nhóm 1 khảo sát một mẫu ngẫu nhiên gồm 50 người và tìm được chiều cao trung bình của 50 người này là 164 cm.
  2. Nhóm 2 khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm.
     Khi đó suy luận về chiều cao trung bình của tổng thể của nhóm 2 đáng tin cậy hơn nhóm 1.
- Một nhóm khảo sát khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm. Khi đó nhóm đưa ra các kết luận
  1. Chiều cao trung bình của tổng thể là 164,6 cm.
  2. Chiều cao trung bình của tổng thể là từ 164 cm đến 165 cm.
  3. Chiều cao trung bình của tổng thể là từ 163,5 cm đến 165,5 cm.
  4. Chiều cao trung bình của tổng thể là từ 130 cm đến 185 cm.

  Trong 3 suy luận đầu tiên, suy luận nào **đáng tin cậy hơn**? (tức là có thể lấy làm kết quả chung cho toàn thành phố)

Định nghĩa 5.1.4

1. **Thống kê mô tả** (Descriptive Statistics) là các phương pháp sử dụng để tóm tắt hoặc mô tả một tập hợp dữ liệu, một mẫu nghiên cứu dưới dạng số hay biểu đồ trực quan.

Định nghĩa 5.1.4

1. **Thống kê mô tả** (Descriptive Statistics) là các phương pháp sử dụng để tóm tắt hoặc mô tả một tập hợp dữ liệu, một mẫu nghiên cứu dưới dạng số hay biểu đồ trực quan.
2. **Thống kê suy luận** (Inferential statistics) bao gồm các phương pháp được sử dụng để suy luận về các đặc điểm tổng thể từ thông tin có trong một mẫu được lấy từ tổng thể này.

5.1.2 Biểu diễn mẫu

5.1.2 Biểu diễn mẫu

Ta có một số cách biểu biễn mẫu như sau:

- Mẫu dạng điểm

5.1.2 Biểu diễn mẫu

Ta có một số cách biểu biễn mẫu như sau:

- Mẫu dạng điểm
- Mẫu dạng tần số (tần suất)

5.1.2 Biểu diễn mẫu

Ta có một số cách biểu biễn mẫu như sau:

- Mẫu dạng điểm
- Mẫu dạng tần số (tần suất)
- Mẫu dạng khoảng

5.1.2 Biểu diễn mẫu

Ta có một số cách biểu biễn mẫu như sau:

- Mẫu dạng điểm
- Mẫu dạng tần số (tần suất)
- Mẫu dạng khoảng
- Mẫu dạng biểu đồ

Định nghĩa 5.1.5

Cho một mẫu có kích thước n, các giá trị của dấu hiệu X mà ta muốn nghiên cứu là $x_1 < x_2 < \ldots < x_m$.

Định nghĩa 5.1.5

Cho một mẫu có kích thước n, các giá trị của dấu hiệu X mà ta muốn nghiên cứu là $x_1 < x_2 < \ldots < x_m$.

1. Số lần lặp lại k_i của x_i được gọi là tần số của x_i. Ta có *Bảng phân bố tần số*

X	x_1	x_2	$\cdots$	x_m
Tần số	k_1	k_2	$\cdots$	k_m

Định nghĩa 5.1.5

Cho một mẫu có kích thước n, các giá trị của dấu hiệu X mà ta muốn nghiên cứu là $x_1 < x_2 < \ldots < x_m$.

1. Số lần lặp lại k_i của x_i được gọi là tần số của x_i. Ta có *Bảng phân bố tần số*

X	x_1	x_2	$\cdots$	x_m
Tần số	k_1	k_2	$\cdots$	k_m

2. Giá trị $f_i = \dfrac{k_i}{n}$ gọi là tần suất của giá trị $x_i, i = 1, 2, \ldots, m$. Ta có *Bảng phân bố tần suất*

X	x_1	x_2	$\cdots$	x_m
Tần suất	f_1	f_2	$\cdots$	f_m

Ví dụ 5.1.6

Kiểm tra 80 hộp, mỗi hộp chứa 100 chip bán dẫn, để tìm số lượng chip bị lỗi trong mỗi hộp. Ta có số chip lỗi trong mỗi hộp cho bởi bảng sau:

Ví dụ 5.1.6

Kiểm tra 80 hộp, mỗi hộp chứa 100 chip bán dẫn, để tìm số lượng chip bị lỗi trong mỗi hộp. Ta có số chip lỗi trong mỗi hộp cho bởi bảng sau:

1	3	4	7	2	7	5	5	2	2	4	2	4	3	2
2	7	1	3	3	2	5	0	0	1	2	5	5	4	1
3	2	6	3	8	2	2	3	1	6	3	4	1	2	5
1	3	3	3	2	1	2	5	5	4	1	4	3	1	0
2	1	2	4	4	5	3	3	4	0	5	2	5	6	2
5	3	3	3	1										

Ta có bảng số liệu sau

Ta có bảng số liệu sau

Số chip bị lỗi	Tần số	Tần suất
0	4	0,05
1	12	0,15
2	18	0,225
3	17	0,2125
4	10	0,125
5	12	0,15
6	3	0,0375
7	3	0,0375
8	1	0,0125
≥ 9	0	0
Tổng	80	1

Khi số các giá trị mà X nhận quá lớn hoặc không đếm được, người ta thường xác định một số khoảng $C_1, C_2, \ldots, C_m$ sao cho mỗi giá trị mà X nhận được chỉ thuộc một khoảng nào đó. Các khoảng này được gọi là **các lớp ghép** của X.

Khi số các giá trị mà X nhận quá lớn hoặc không đếm được, người ta thường xác định một số khoảng $C_1, C_2, \ldots, C_m$ sao cho mỗi giá trị mà X nhận được chỉ thuộc một khoảng nào đó. Các khoảng này được gọi là **các lớp ghép** của X.

Ví dụ 5.1.7

Một mẫu về chiều cao của 40 sinh viên được trình bày trong bảng phân bố ghép lớp sau:

Khoảng	Tần số	Tần suất
$(146; 151]$	4	0,1
$(151; 156]$	2	0,05
$(156; 161]$	6	0,15
$(161; 166]$	10	0,25
$(166; 171]$	12	0,3
$(171; 176]$	6	0,15

Trong một mẫu, dấu hiệu điều tra X có bảng phân bố tần số và tần suất

x_i	x_1	x_2	$\cdots$	x_m
Tần số	k_1	k_2	$\cdots$	k_m
Tần suất	f_1	f_2	$\cdots$	f_m

Trên mặt phẳng toạ độ, nối điểm $(x_i; k_i)$ với điểm $(x_i; 0)$ bởi đoạn thẳng, với $i = 1; \ldots, m$ ta được **biểu đồ tần số hình gậy**.

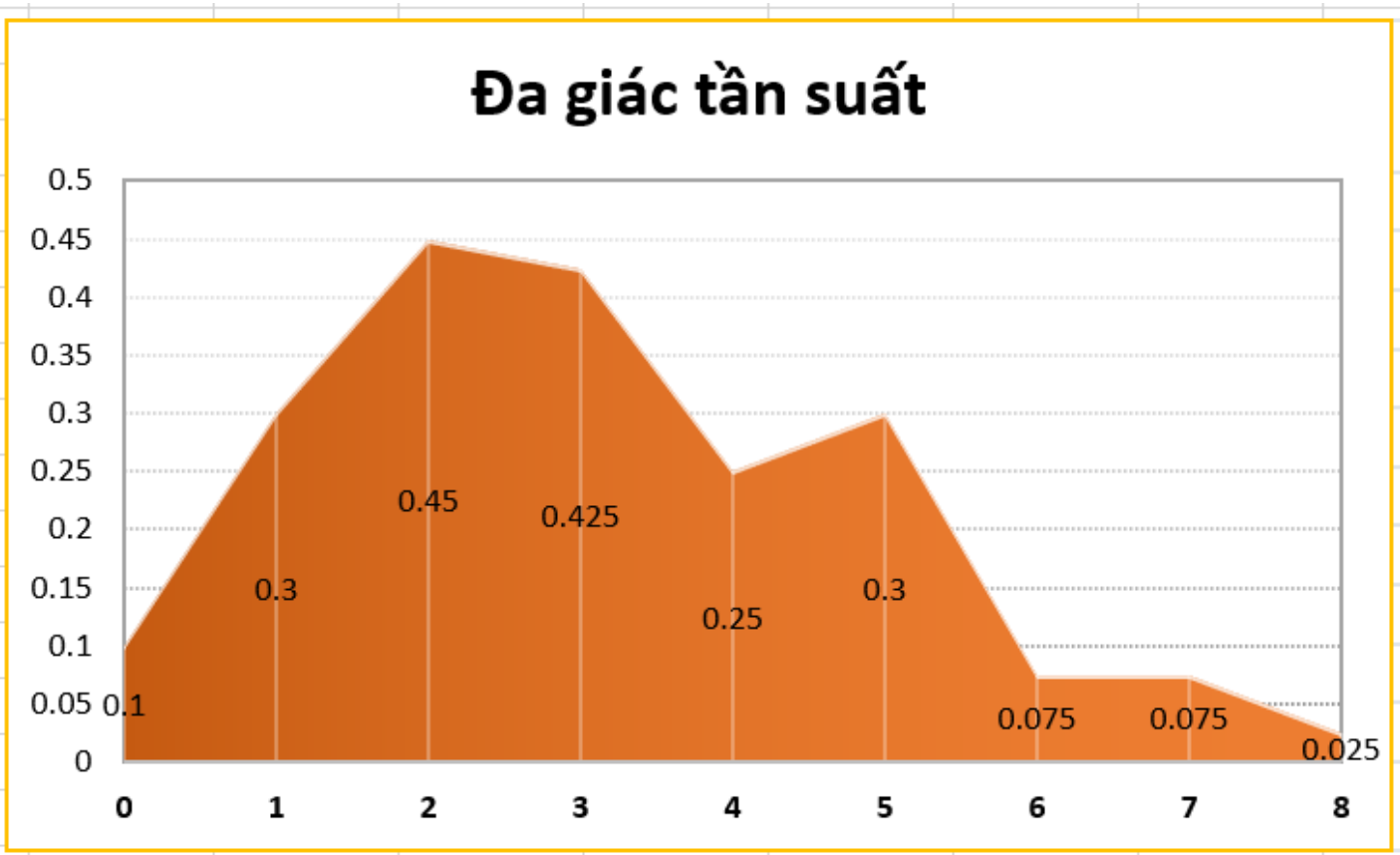
Trong một mẫu, dấu hiệu điều tra X có bảng phân bố tần số và tần suất

x_i	x_1	x_2	$\cdots$	x_m
Tần số	k_1	k_2	$\cdots$	k_m
Tần suất	f_1	f_2	$\cdots$	f_m

Trên mặt phẳng toạ độ, nối điểm $(x_i; k_i)$ với điểm $(x_i; 0)$ bởi đoạn thẳng, với $i = 1; \ldots, m$ ta được **biểu đồ tần số hình gậy**.
Biểu đồ tần số hình gậy của Ví dụ 5.1.6:

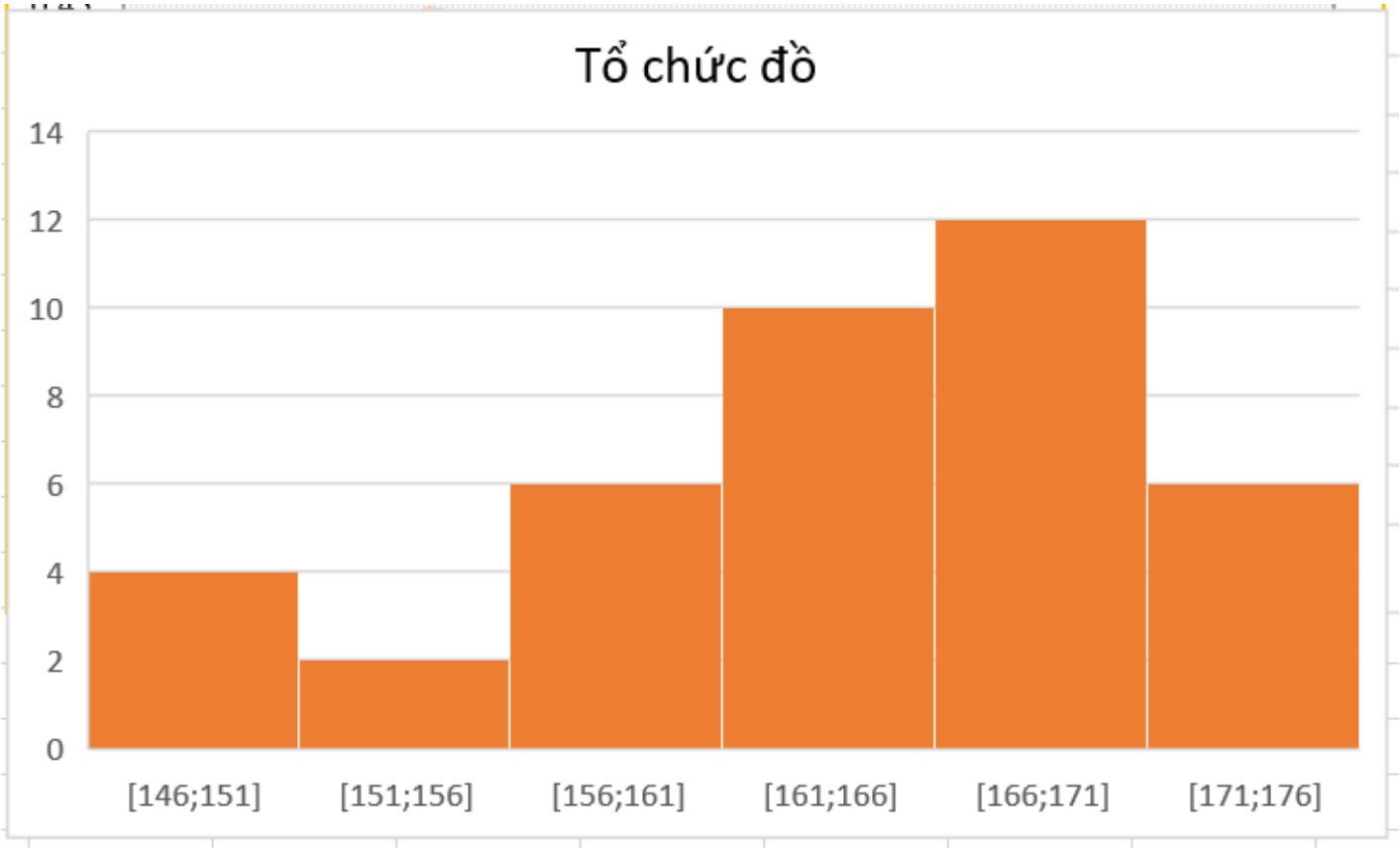
Nối các điểm $(x_i; f_i)$ với $(x_{i+1}; f_{i+1})$ bởi đoạn thẳng, với $i = 1, 2, \ldots k-1$ ta được đường gấp khúc được gọi là **biểu đồ đa giác tần suất**.

Nối các điểm $(x_i; f_i)$ với $(x_{i+1}; f_{i+1})$ bởi đoạn thẳng, với $i = 1, 2, \ldots k-1$ ta được đường gấp khúc được gọi là **biểu đồ đa giác tần suất**.
Biểu đồ đa giác tần suất của Ví dụ 5.1.6:



Đối với bản phân bố lớp ghép, người ta thường dùng **tổ chức đồ** (histogram) để biểu diễn.

Đối với bản phân bố lớp ghép, người ta thường dùng **tổ chức đồ** (histogram) để biểu diễn.
Ta có tổ chức đồ tần số của Ví dụ 5.1.7:



Người ta có thể dùng các biểu đồ khác nhau để biểu diễn dữ liệu thu được.

5.1.3 Các tham số của mẫu

5.1.3 Các tham số của mẫu

Cho một mẫu ngẫu nhiên có kích thước n được lấy từ một tổng thể. Các quan sát từ mẫu là các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$.

5.1.3 Các tham số của mẫu

Cho một mẫu ngẫu nhiên có kích thước n được lấy từ một tổng thể. Các quan sát từ mẫu là các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$.

Định nghĩa 5.1.8 (Trung bình mẫu)

1. Trung bình mẫu ngẫu nhiên $\{X_1, \ldots, X_n\}$ được xác định bởi

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

5.1.3 Các tham số của mẫu

Cho một mẫu ngẫu nhiên có kích thước n được lấy từ một tổng thể. Các quan sát từ mẫu là các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$.

Định nghĩa 5.1.8 (Trung bình mẫu)

1. Trung bình mẫu ngẫu nhiên $\{X_1, \ldots, X_n\}$ được xác định bởi

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

2. Nếu một mẫu có kích thước n nhận các giá trị $x_1, \ldots, x_n$ thì **trung bình mẫu cụ thể** là

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

Nếu có một phân bố ghép lớp với m khoảng $C_1, C_2, \ldots, C_m$ và tần số của khoảng C_i là k_i thì trung bình mẫu cụ thể $\overline{x}$ được xác định bởi

$$\overline{x} = \frac{\sum_{i=1}^{m} k_i x_i}{\sum_{i=1}^{m} k_i}$$

trong đó x_i là trung điểm (tâm) của khoảng C_i.

Nếu có một phân bố ghép lớp với m khoảng $C_1, C_2, \ldots, C_m$ và tần số của khoảng C_i là k_i thì trung bình mẫu cụ thể $\overline{x}$ được xác định bởi

$$\overline{x} = \frac{\sum_{i=1}^{m} k_i x_i}{\sum_{i=1}^{m} k_i}$$

trong đó x_i là trung điểm (tâm) của khoảng C_i.

Định nghĩa 5.1.9 (Trung vị mẫu)

Trung vị mẫu, kí hiệu m, là một số mà số các giá trị của mẫu $\geq m$ bằng số các giá trị của mẫu $\leq m$.

Nếu có một phân bố ghép lớp với m khoảng $C_1, C_2, \ldots, C_m$ và tần số của khoảng C_i là k_i thì trung bình mẫu cụ thể $\overline{x}$ được xác định bởi

$$\overline{x} = \frac{\sum_{i=1}^{m} k_i x_i}{\sum_{i=1}^{m} k_i}$$

trong đó x_i là trung điểm (tâm) của khoảng C_i.

Định nghĩa 5.1.9 (Trung vị mẫu)

Trung vị mẫu, kí hiệu m, là một số mà số các giá trị của mẫu $\geq m$ bằng số các giá trị của mẫu $\leq m$.

Định nghĩa 5.1.10 (Mode mẫu)

Mode của mẫu là giá trị của mẫu có tần số lớn nhất.

Định nghĩa 5.1.11 (Phương sai)

1. Một tổng thể kích thước N có trung bình tổng thể là μ. Phương sai tổng thể (variance of a population), ký hiệu σ^2, xác định bởi

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}.$$

Định nghĩa 5.1.11 (Phương sai)

1. Một tổng thể kích thước N có trung bình tổng thể là μ. Phương sai tổng thể (variance of a population), ký hiệu σ^2, xác định bởi

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}.$$

2. Cho một mẫu có kích thước n và trung bình mẫu là $\overline{x}$. Phương sai mẫu chưa hiệu chỉnh (variance of a sample), ký hiệu $\hat{s}^2$, xác định bởi

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}.$$

Định nghĩa 5.1.11 (Phương sai)

1. Một tổng thể kích thước N có trung bình tổng thể là μ. Phương sai tổng thể (variance of a population), ký hiệu σ^2, xác định bởi

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}.$$

2. Cho một mẫu có kích thước n và trung bình mẫu là $\overline{x}$. Phương sai mẫu chưa hiệu chỉnh (variance of a sample), ký hiệu $\hat{s}^2$, xác định bởi

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}.$$

3. **Phương sai mẫu hiệu chỉnh**: $s^2 = \frac{n}{n-1}\hat{s}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$.

Định nghĩa 5.1.11 (Phương sai)

1. Một tổng thể kích thước N có trung bình tổng thể là μ. Phương sai tổng thể (variance of a population), ký hiệu σ^2, xác định bởi

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}.$$

2. Cho một mẫu có kích thước n và trung bình mẫu là $\overline{x}$. Phương sai mẫu chưa hiệu chỉnh (variance of a sample), ký hiệu $\hat{s}^2$, xác định bởi

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}.$$

3. **Phương sai mẫu hiệu chỉnh**: $s^2 = \frac{n}{n-1}\hat{s}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$.
   s được gọi là độ lệch chuẩn mẫu hiệu chỉnh.

Định nghĩa 5.1.12

Một phân phối của trung bình mẫu (a sampling distribution of sample means) là một phân phối xác suất của các trung bình mẫu lấy được từ các mẫu ngẫu nhiên cùng kích thước được lấy từ một tổng thể.

Định nghĩa 5.1.12

Một phân phối của trung bình mẫu (a sampling distribution of sample means) là một phân phối xác suất của các trung bình mẫu lấy được từ các mẫu ngẫu nhiên cùng kích thước được lấy từ một tổng thể.

Ví dụ 5.1.13 Cho một tổng thể có kích thước $N = 5$ gồm các số $3, 6, 9, 12, 15$. Chọn một mẫu có kích thước $n = 3$ (chọn không lặp lại). Tìm phân phối của trung bình mẫu $\overline{x}$.

Giải. Ta có 10 mẫu khác nhau có kích thước $n = 3$.

Mẫu	Các giá trị của mẫu	$\overline{x}$
1	3, 6, 9	6
2	3, 6, 12	7
3	3, 6, 15	8
4	3, 9, 12	8
5	3, 9, 15	9
6	3, 12, 15	10
7	6, 9, 12	9
8	6, 9, 15	10
9	6, 12, 15	11
10	9, 12, 15	12

Mẫu	Các giá trị của mẫu	$\overline{x}$
1	3, 6, 9	6
2	3, 6, 12	7
3	3, 6, 15	8
4	3, 9, 12	8
5	3, 9, 15	9
6	3, 12, 15	10
7	6, 9, 12	9
8	6, 9, 15	10
9	6, 12, 15	11
10	9, 12, 15	12

Bảng phân phối của trung bình mẫu

$\overline{X}$	6	7	8	9	10	11	12
$P(\overline{X} = x_i)$	0,1	0,1	0,2	0,2	0,2	0,1	0,1

Nhận xét 5.1.14 Theo Định lý giới hạn trung tâm:

1. Nếu tổng thể có phân phối chuẩn $N(\mu;\sigma^2)$ thì $\overline{X} \sim N(\mu;\sigma^2/n)$ và $S_n \sim N(n\mu;n\sigma^2)$ với mọi kích thước mẫu n.

Nhận xét 5.1.14 Theo Định lý giới hạn trung tâm:

1. Nếu tổng thể có phân phối chuẩn $N(\mu;\sigma^2)$ thì $\overline{X} \sim N(\mu;\sigma^2/n)$ và $S_n \sim N(n\mu;n\sigma^2)$ với mọi kích thước mẫu n.
2. Nếu tổng thể có phân phối không chuẩn thì $\overline{X} \approx N(\mu;\dfrac{\sigma^2}{n})$ và $S_n = \sum_1^n X_n \approx N(n\mu;n\sigma^2)$ với $n \geq 30$.

Ví dụ 5.1.15

Giả sử chiều cao của sinh viên nam ở TPHCM có phân phối chuẩn với trung bình là 172cm và độ lệch chuẩn là 10cm. Chọn một mẫu ngẫu nhiên gồm 25 sinh viên.

a. Tìm quy luật phân phối của trung bình mẫu.

b. Tính xác suất mẫu đó có chiều cao trung bình lớn hơn 174cm.

Ví dụ 5.1.15

Giả sử chiều cao của sinh viên nam ở TPHCM có phân phối chuẩn với trung bình là 172cm và độ lệch chuẩn là 10cm. Chọn một mẫu ngẫu nhiên gồm 25 sinh viên.
a. Tìm quy luật phân phối của trung bình mẫu.
b. Tính xác suất mẫu đó có chiều cao trung bình lớn hơn 174cm.

Giải. a. Gọi X là chiều cao của sinh viên TPHCM (tổng thể). Đặt X_i là chiều cao của sinh viên i $(i = 1, 2, \ldots, 25)$ và $\overline{X}$ là chiều cao trung bình của mẫu ngẫu nhiên gồm 25 sinh viên.

Ví dụ 5.1.15

Giả sử chiều cao của sinh viên nam ở TPHCM có phân phối chuẩn với trung bình là 172cm và độ lệch chuẩn là 10cm. Chọn một mẫu ngẫu nhiên gồm 25 sinh viên.
a. Tìm quy luật phân phối của trung bình mẫu.
b. Tính xác suất mẫu đó có chiều cao trung bình lớn hơn 174cm.

Giải. a. Gọi X là chiều cao của sinh viên TPHCM (tổng thể). Đặt X_i là chiều cao của sinh viên i $(i = 1, 2, \ldots, 25)$ và $\overline{X}$ là chiều cao trung bình của mẫu ngẫu nhiên gồm 25 sinh viên.
Vì $X \sim N(\mu; \sigma^2)$ nên theo Định lý giới hạn trung tâm, ta có
$\overline{X} \sim N(\mu, \sigma^2/n)$, trong đó $\mu = 172; \sigma^2/n = \dfrac{10^2}{25} = 4.$

Ví dụ 5.1.15

Giả sử chiều cao của sinh viên nam ở TPHCM có phân phối chuẩn với trung bình là 172cm và độ lệch chuẩn là 10cm. Chọn một mẫu ngẫu nhiên gồm 25 sinh viên.

a. Tìm quy luật phân phối của trung bình mẫu.

b. Tính xác suất mẫu đó có chiều cao trung bình lớn hơn 174cm.

Giải. a. Gọi X là chiều cao của sinh viên TPHCM (tổng thể). Đặt X_i là chiều cao của sinh viên i $(i = 1, 2, \ldots, 25)$ và $\overline{X}$ là chiều cao trung bình của mẫu ngẫu nhiên gồm 25 sinh viên.

Vì $X \sim N(\mu; \sigma^2)$ nên theo Định lý giới hạn trung tâm, ta có

$\overline{X} \sim N(\mu, \sigma^2/n)$, trong đó $\mu = 172; \sigma^2/n = \dfrac{10^2}{25} = 4.$

b. Ta có

$$P(\overline{X} > 174) = 1 - P(\overline{X} \leq 174) = 1 - \Phi(\frac{174 - 172}{2})$$
$$= 1 - \Phi(1) = 1 - 0,8413 = 0,1587.$$

- Đặt p là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong tổng thể;

- Đặt p là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong tổng thể;
- Đặt F là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong mẫu;

- Đặt p là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong tổng thể;
- Đặt F là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong mẫu;
- Khi $np > 5$ và $n(1-p) > 5$, ta có $F \approx N(p; \dfrac{p(1-p)}{n})$

- Đặt p là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong tổng thể;
- Đặt F là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong mẫu;
- Khi $np > 5$ và $n(1-p) > 5$, ta có $F \approx N(p; \dfrac{p(1-p)}{n})$

Ví dụ 5.1.16

Người ta phát hiện ra một nhà máy sản xuất có 2% sản phẩm do nhà máy sản xuất có lỗi. Tính xác suất trong 400 sản phẩm do máy này sản xuất có không dưới 3% sản phẩm bị lỗi.

- Đặt p là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong tổng thể;
- Đặt F là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong mẫu;
- Khi $np > 5$ và $n(1-p) > 5$, ta có $F \approx N(p; \dfrac{p(1-p)}{n})$

Ví dụ 5.1.16

Người ta phát hiện ra một nhà máy sản xuất có 2% sản phẩm do nhà máy sản xuất có lỗi. Tính xác suất trong 400 sản phẩm do máy này sản xuất có không dưới 3% sản phẩm bị lỗi.

Giải. Đặt F là tỉ lệ sản phẩm có lỗi của mẫu. Tỉ lệ tổng thể là $p = 0,02$. Vì kích thước mẫu $n = 400$ nên tỉ lệ sản phẩm lỗi của mẫu $F \approx N(p; \dfrac{p(1-p)}{n})$ trong đó $p = 0,02$ và $\sqrt{\dfrac{p(1-p)}{n}} = 0,007$.

- Đặt p là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong tổng thể;
- Đặt F là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong mẫu;
- Khi $np > 5$ và $n(1-p) > 5$, ta có $F \approx N(p; \frac{p(1-p)}{n})$

Ví dụ 5.1.16

Người ta phát hiện ra một nhà máy sản xuất có 2% sản phẩm do nhà máy sản xuất có lỗi. Tính xác suất trong 400 sản phẩm do máy này sản xuất có không dưới 3% sản phẩm bị lỗi.

Giải. Đặt F là tỉ lệ sản phẩm có lỗi của mẫu. Tỉ lệ tổng thể là $p = 0,02$. Vì kích thước mẫu $n = 400$ nên tỉ lệ sản phẩm lỗi của mẫu $F \approx N(p; \frac{p(1-p)}{n})$ trong đó $p = 0,02$ và $\sqrt{\frac{p(1-p)}{n}} = 0,007$. Do đó

$$P(F \geq 0,03) = 1 - P(F < 0,03) = 1 - \Phi_F(\frac{0,03 - 0,02}{0,007})$$

$$= 1 - \Phi(1,43) = 1 - 0,9236 = 0,0764.$$

Như vậy, xác suất mẫu có không dưới 3% sản phẩm bị lỗi là 7,64%.

5.2 Lý thuyết ước lượng

Ước lượng là một phỏng đoán về một giá trị chưa biết của tổng thể dựa vào quan sát trên mẫu lấy ra từ tổng thể.

Ước lượng là một phỏng đoán về một giá trị chưa biết của tổng thể dựa vào quan sát trên mẫu lấy ra từ tổng thể.

Định nghĩa 5.2.1

1. Một ước lượng điểm là một giá trị dùng để ước lượng một tham số.
2. Một ước lượng khoảng là một khoảng giá trị dùng để ước lượng một tham số.

Ước lượng là một phỏng đoán về một giá trị chưa biết của tổng thể dựa vào quan sát trên mẫu lấy ra từ tổng thể.

Định nghĩa 5.2.1

1. Một ước lượng điểm là một giá trị dùng để ước lượng một tham số.
2. Một ước lượng khoảng là một khoảng giá trị dùng để ước lượng một tham số.

Ví dụ.

- Nếu nói chiều cao trung bình của sinh viên nam trường Đại học Công nghệ Thông tin là 174 cm thì đó là một **ước lượng điểm**.
- Nếu nói chiều cao trung bình đó nằm trong khoảng từ 159 cm đến 169 cm hay 164 ± 5 cm. Khi đó ta đã có một **ước lượng khoảng.**

Các bài toán về ước lượng:

Các bài toán về ước lượng:

1. Ước lượng khoảng cho trung bình tổng thể khi biết độ lệch chuẩn tổng thể và trung bình mẫu.

Các bài toán về ước lượng:

1. Ước lượng khoảng cho trung bình tổng thể khi biết độ lệch chuẩn tổng thể và trung bình mẫu.
2. Ước lượng khoảng cho trung bình tổng thể khi chưa biết độ lệch chuẩn tổng thể.

Các bài toán về ước lượng:

1. Ước lượng khoảng cho trung bình tổng thể khi biết độ lệch chuẩn tổng thể và trung bình mẫu.
2. Ước lượng khoảng cho trung bình tổng thể khi chưa biết độ lệch chuẩn tổng thể.
3. Ước lượng tỉ lệ của tổng thể.

5.2.1 Ước lượng khoảng cho trung bình tổng thể khi biết độ lệch chuẩn tổng thể và trung bình mẫu

5.2.1 Ước lượng khoảng cho trung bình tổng thể khi biết độ lệch chuẩn tổng thể và trung bình mẫu

Bài toán 1.

Giả sử rằng thời gian mua sắm của khách hàng tại một trung tâm thương mại có phân phối chuẩn với độ lệch chuẩn tổng thể là 20 phút. Chọn ngẫu nhiên 64 người đã mua sắm ở trung tâm đó. Người ta thấy rằng thời gian mua sắm trung bình của 64 người này là 75 phút. Tìm thời gian mua sắm trung bình của khách hàng tại trung tâm này với độ tin cậy 95%.

Định nghĩa 5.2.2

1. **Độ tin cậy** (confidence level), ký hiệu $1-\alpha$, của ước lượng khoảng của một tham số là xác suất khoảng ước lượng chứa tham số. Giả sử một số lượng lớn mẫu được chọn và quá trình ước lượng trên cùng một tham số được lặp lại.

Định nghĩa 5.2.2

1. **Độ tin cậy** (confidence level), ký hiệu $1-\alpha$, của ước lượng khoảng của một tham số là xác suất khoảng ước lượng chứa tham số. Giả sử một số lượng lớn mẫu được chọn và quá trình ước lượng trên cùng một tham số được lặp lại.
2. **Khoảng tin cậy** (confidence interval) là một khoảng ước lượng cụ thể của một tham số tương ứng với độ tin cậy đã cho.

Định nghĩa 5.2.2

1. **Độ tin cậy** (confidence level), ký hiệu $1-\alpha$, của ước lượng khoảng của một tham số là xác suất khoảng ước lượng chứa tham số. Giả sử một số lượng lớn mẫu được chọn và quá trình ước lượng trên cùng một tham số được lặp lại.
2. **Khoảng tin cậy** (confidence interval) là một khoảng ước lượng cụ thể của một tham số tương ứng với độ tin cậy đã cho.

Ví dụ: Khi nói **khoảng tin cậy** của chiều cao trung bình của sinh viên các trường đại học tại TPHCM là $[155; 175]$ với **độ tin cậy** 95% có nghĩa là xác suất khoảng $[155; 175]$ chứa trung bình tổng thể là 95%.

Bài toán: Ước lượng trung bình tổng thể μ khi biết σ và độ tin cậy $1-\alpha$.

Bài toán: Ước lượng trung bình tổng thể μ khi biết σ và độ tin cậy $1-\alpha$.

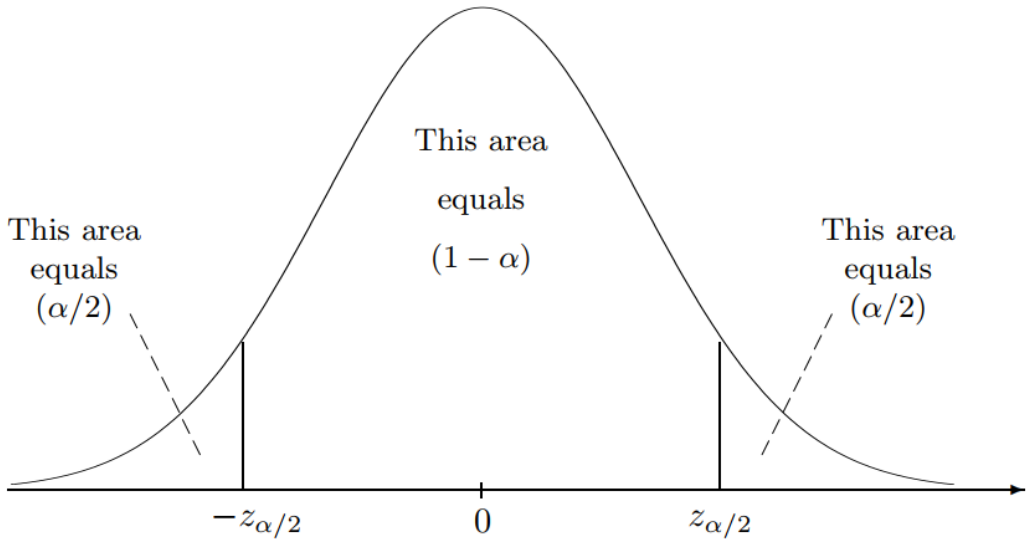
- Nếu tổng thể có phân phối chuẩn hoặc tổng thể có phân phối không chuẩn và kích thước mẫu $n \geq 30$ thì $\overline{X}$ sẽ xấp xỉ phân phối chuẩn, $\overline{X} \approx N(\mu; \frac{\sigma^2}{n})$.

Bài toán: Ước lượng trung bình tổng thể μ khi biết σ và độ tin cậy $1-\alpha$.

- Nếu tổng thể có phân phối chuẩn hoặc tổng thể có phân phối không chuẩn và kích thước mẫu $n \geq 30$ thì $\overline{X}$ sẽ xấp xỉ phân phối chuẩn, $\overline{X} \approx N(\mu; \frac{\sigma^2}{n})$.
- Đổi biến

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}},$$

khi đó $Z \sim N(0;1)$.



Với độ tin cậy $1-\alpha$ thì giá trị $z_{\alpha/2}$ xác định từ

$$\boxed{\Phi(z_{\alpha/2}) = 1 - \alpha/2}.$$

Với độ tin cậy $1-\alpha$ thì giá trị $z_{\alpha/2}$ xác định từ

$$\boxed{\Phi(z_{\alpha/2}) = 1 - \alpha/2}.$$

Tra bảng ta tìm được $z_{\alpha/2}$. Do đó

$$-z_{\alpha/2} \leq \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}$$

Với độ tin cậy $1-\alpha$ thì giá trị $z_{\alpha/2}$ xác định từ

$$\boxed{\Phi(z_{\alpha/2}) = 1 - \alpha/2}.$$

Tra bảng ta tìm được $z_{\alpha/2}$. Do đó

$$-z_{\alpha/2} \leq \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}$$

Trung bình tổng thể sẽ thuộc khoảng (khoảng tin cậy của trung bình tổng thể với độ tin cậy $1-\alpha$)

$$\boxed{[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]}$$

Với độ tin cậy $1-\alpha$ thì giá trị $z_{\alpha/2}$ xác định từ

$$\boxed{\Phi(z_{\alpha/2}) = 1 - \alpha/2}.$$

Tra bảng ta tìm được $z_{\alpha/2}$. Do đó

$$-z_{\alpha/2} \leq \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}$$

Trung bình tổng thể sẽ thuộc khoảng (khoảng tin cậy của trung bình tổng thể với độ tin cậy $1-\alpha$)

$$\boxed{[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]}$$

$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ được gọi là **sai số của ước lượng** hoặc **độ chính xác** của ước lượng.

Giải Bài toán 1. Giả sử rằng thời gian mua sắm của khách hàng tại một trung tâm thương mại có phân phối chuẩn với độ lệch chuẩn tổng thể là 20 phút. Chọn ngẫu nhiên 64 người đã mua sắm ở trung tâm đó. Người ta thấy rằng thời gian mua sắm trung bình của 64 người này là 75 phút. Tìm thời gian mua sắm trung bình của khách hàng tại trung tâm này với độ tin cậy 95%.

- Ta có $\overline{x} = ...; n = ...$ và $\sigma = ...;$
- Độ tin cậy $1 - \alpha = 95\%$. Suy ra $\alpha = ...$ và $z_{\alpha/2} = ...$
- Độ chính xác $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} = ...\dfrac{.....}{\sqrt{.....}} = ...$
- Khoảng tin cậy của trung bình tổng thể với độ tin cậy 95% là

$$[.....].$$

Bài tập 1. Chọn ngẫu nhiên 30 người để kiểm tra thời gian sử dụng chiếc điện thoại di động đầu tiên. Người ta thấy rằng thời gian sử dụng trung bình của 30 người này là 5,6 năm. Giả sử thời gian sử dụng chiếc điện thoại di động đầu tiên có phân phối chuẩn với độ lệch chuẩn là 0,8 năm. Tính khoảng thời gian trung bình sử dụng chiếc điện thoại đầu tiên với độ tin cậy 99%.

Bài tập 2. Thu nhập trung bình hàng tháng của 30 hộ dân trong một thành phố được cho như sau (đơn vị triệu đồng)

12.23	16.56	4.39	2.89	1.24	2.17	13.19	9.16	1.42	73.25
1.91	14.64	11.59	6.69	1.06	8.74	3.17	18.13	7.92	4.78
16.85	40.22	2.42	21.58	5.01	1.47	12.24	2.27	12.77	2.76

Tìm khoảng tin cậy 90% của thu nhập trung bình hàng tháng của toàn thành phố. Biết thu nhập trung bình hàng tháng có phân phối chuẩn và có độ lệch chuẩn 14.405.

5.2.2 Ước lượng khoảng cho trung bình tổng thể khi chưa biết độ lệch chuẩn tổng thể

5.2.2 Ước lượng khoảng cho trung bình tổng thể khi chưa biết độ lệch chuẩn tổng thể

Bài toán 2. Thu nhập trung bình hàng tháng của **30** hộ dân trong một thành phố được cho như sau (đơn vị triệu đồng)

12.23	16.56	4.39	2.89	1.24	2.17	13.19	9.16	1.42	73.25
1.91	14.64	11.59	6.69	1.06	8.74	3.17	18.13	7.92	4.78
16.85	40.22	2.42	21.58	5.01	1.47	12.24	2.27	12.77	2.76

Tìm khoảng thu nhập trung bình hàng tháng của toàn thành phố với độ tin cậy 90%.

Bài toán 3. Theo một thống kê cho thấy số thu nhập của 7 công nhân trong năm 2021 của một công ty được cho như sau (đơn vị triệu đồng)

54,6 59 60,9 63,1 71,6 84,4 99,3

Giả sử thu nhập trong năm 2021 của công ty có phân phối chuẩn. Tính khoảng thu nhập trung bình của công ty này với độ tin cậy 99%.

5.2.2 Ước lượng khoảng cho trung bình tổng thể khi chưa biết độ lệch chuẩn tổng thể

- Cho một tổng thể có phân phối chuẩn với trung bình μ.
- Cho $\overline{X}$ là trung bình của mẫu ngẫu nhiên và $\overline{x}, s$ lần lượt là trung bình và độ lệch chuẩn hiệu chỉnh của một mẫu có kích thước n.
- Biến ngẫu nhiên

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

  có phân phối Student với bậc tự do $n - 1$.
- khoảng tin cậy của μ với độ tin cậy $(1 - \alpha)$ là

$$\left[\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right].$$

5.2.2 Ước lượng khoảng cho trung bình tổng thể khi chưa biết độ lệch chuẩn tổng thể

Trường hợp 1: Kích thước mẫu $n \geq 30$.

5.2.2 Ước lượng khoảng cho trung bình tổng thể khi chưa biết độ lệch chuẩn tổng thể

Trường hợp 1: Kích thước mẫu $n \geq 30$.

1. $\overline{x}, s$ là trung bình và độ lệch chuẩn hiệu chỉnh của một mẫu cụ thể

5.2.2 Ước lượng khoảng cho trung bình tổng thể khi chưa biết độ lệch chuẩn tổng thể

Trường hợp 1: Kích thước mẫu $n \geq 30$.

1. $\overline{x}, s$ là trung bình và độ lệch chuẩn hiệu chỉnh của một mẫu cụ thể
2. Đổi biến $Z = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$, khi đó $Z \sim N(0;1)$.

5.2.2 Ước lượng khoảng cho trung bình tổng thể khi chưa biết độ lệch chuẩn tổng thể

Trường hợp 1: Kích thước mẫu $n \geq 30$.

1. $\overline{x}, s$ là trung bình và độ lệch chuẩn hiệu chỉnh của một mẫu cụ thể
2. Đổi biến $Z = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$, khi đó $Z \sim N(0;1)$.
3. Tra bảng $A4$, tìm $z_{\alpha/2}$.

5.2.2 Ước lượng khoảng cho trung bình tổng thể khi chưa biết độ lệch chuẩn tổng thể

Trường hợp 1: Kích thước mẫu $n \geq 30$.

1. $\overline{x}, s$ là trung bình và độ lệch chuẩn hiệu chỉnh của một mẫu cụ thể
2. Đổi biến $Z = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$, khi đó $Z \sim N(0;1)$.
3. Tra bảng $A4$, tìm $z_{\alpha/2}$.
4. Khoảng tin cậy của μ với độ tin cậy $1-\alpha$ là

$$[\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}].$$

Trường hợp 2: Kích thước mẫu $n < 30$ và tổng thể có phân phối chuẩn

Trường hợp 2: Kích thước mẫu $n < 30$ và tổng thể có phân phối chuẩn

1. Đổi biến $T = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$, khi đó $T \sim St(n)$.

Trường hợp 2: Kích thước mẫu $n < 30$ và tổng thể có phân phối chuẩn

1. Đổi biến $T = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$, khi đó $T \sim St(n)$.
2. Tra bảng $A5$ dòng $n-1$, tìm $t_{\alpha/2}$ thỏa mãn $P(T > t_{\alpha/2}) = \dfrac{\alpha}{2}$.

Trường hợp 2: Kích thước mẫu $n < 30$ và tổng thể có phân phối chuẩn

1. Đổi biến $T = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$, khi đó $T \sim St(n)$.
2. Tra bảng $A5$ dòng $n-1$, tìm $t_{\alpha/2}$ thỏa mãn $P(T > t_{\alpha/2}) = \dfrac{\alpha}{2}$.
3. Khoảng tin cậy của μ với độ tin cậy $1-\alpha$ là

$$[\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}].$$

Cách tìm t_β trong bảng A5: bậc tự do của phân phối Student là $n-1$. Cột đầu tiên bên trái của bảng A5 là cột bậc tự do, hai hàng đầu tiên bên trên là giá trị của β. Số nằm ở vị trí của giao của hàng tương ứng với bậc tự do $n-1$ và cột tương ứng với β là giá trị của t_β.

Cách tìm t_β trong bảng A5: bậc tự do của phân phối Student là $n-1$. Cột đầu tiên bên trái của bảng A5 là cột bậc tự do, hai hàng đầu tiên bên trên là giá trị của β. Số nằm ở vị trí của giao của hàng tương ứng với bậc tự do $n-1$ và cột tương ứng với β là giá trị của t_β.

Ví dụ. Tìm giá trị $t_{0,005}$ với bậc tự do 17. Theo bảng A5, ta có $t_{0,005} = 2,898$.

ν	α, the right-tail probability									
(d.f.)	.10	.05	.025	.02	.01	.005	.0025	.001	.0005	.0001
1	3.078	6.314	12.706	15.89	31.82	63.66	127.3	318.3	636.6	3185
2	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60	70.71
3	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92	22.20
4	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610	13.04
5	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.894	6.869	9.676
6	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959	8.023
7	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408	7.064
8	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041	6.442
9	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781	6.009
10	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587	5.694
11	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437	5.453
12	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318	5.263
13	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221	5.111
14	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140	4.985
15	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073	4.880
16	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015	4.790
17	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965	4.715
18	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922	4.648
19	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883	4.590
20	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850	4.539

Bài toán 3. Theo một thống kê cho thấy số thu nhập của 7 công nhân trong năm 2021 của một công ty được cho như sau (đơn vị triệu đồng)

54,6 59 60,9 63,1 71,6 84,4 99,3

Giả sử thu nhập trong năm 2021 của công ty có phân phối chuẩn. Tính khoảng thu nhập trung bình của công ty này với độ tin cậy 99%.

Giải Bài toán 3.

- Trung bình mẫu: $\overline{x} =$ .....
- Độ lệch chuẩn mẫu hiệu chỉnh: $s =$ .....
- Tìm $t_{\alpha/2}$ với độ tin cậy $1-\alpha = 0,99$ và bậc tự do 6. Ta có $t_{\alpha/2} =$ .....
- Khoảng tin cậy cần tìm

Bài toán 3. Theo một thống kê cho thấy số thu nhập của 7 công nhân trong năm 2021 của một công ty được cho như sau (đơn vị triệu đồng)

54,6 59 60,9 63,1 71,6 84,4 99,3

Giả sử thu nhập trong năm 2021 của công ty có phân phối chuẩn. Tính khoảng thu nhập trung bình của công ty này với độ tin cậy 99%.

Giải Bài toán 3.

- Trung bình mẫu: $\overline{x} =$ .....
- Độ lệch chuẩn mẫu hiệu chỉnh: $s =$ .....
- Tìm $t_{\alpha/2}$ với độ tin cậy $1-\alpha = 0,99$ và bậc tự do 6. Ta có $t_{\alpha/2} =$ .....
- Khoảng tin cậy cần tìm

$$\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$70,414 - 3,707\frac{16,103}{\sqrt{7}} \leq \mu \leq 70,414 + 3,707\frac{16,103}{\sqrt{7}}$$

$$47,852 \leq \mu \leq 92,976$$

Ví dụ 5.2.4

Kiểm tra tuổi thọ (tính bằng giờ) của 50 bóng đèn do nhà máy A sản xuất, người ta được bảng số liệu sau

Tuổi thọ	3300	3500	3600	4000
Số bóng đèn	10	20	12	8

a. Ước tính tuổi thọ trung bình của các bóng đèn do nhà máy A sản xuất với độ tin cậy 97%.
b. Dựa vào mẫu trên để ước lượng tuổi thọ trung bình của các bóng đèn do nhà máy A sản xuất có độ chính xác 59,02 giờ thì phải đảm bảo độ tin cậy là bao nhiêu?
c. Dựa vào mẫu trên, nếu muốn ước lượng tuổi thọ trung bình của các bóng đèn do nhà máy A sản suất có độ chính xác nhỏ hơn 40 giờ với độ tin cậy 98% thì cần phải kiểm tra tối thiểu bao nhiêu bóng đèn?

Giải a. (Kích thước mẫu bằng 50 và chưa biết độ lệch chuẩn tổng thể)

- Trung bình mẫu: $\overline{x} =$ .....
- Độ lệch chuẩn mẫu hiệu chỉnh: $s =$ .....
- Độ tin cậy $1 - \alpha = 0,97$. Suy ra $\Phi(z_{\alpha/2}) = 1 - \alpha/2 = 0,985$. Do đó $z_{\alpha/2} =$ .....
- Độ chính xác:

$$z_{\alpha/2}\frac{s}{\sqrt{n}} = 2,17\frac{217,3683}{\sqrt{50}} = ...$$

- Khoảng tin cậy của tuổi thọ trung của bóng đèn với độ tin cậy 97% là

.....

b. Ta có độ chính xác bằng ..... giờ, tức là

$$z_{\alpha/2}\frac{s}{\sqrt{n}} = ...$$

Suy ra

$$z_{\alpha/2} = 59,02\frac{\sqrt{n}}{s} = ...$$

Do đó

$$\Phi(1,92) = \Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}.$$

Trang bảng A4, ta có $\Phi(.....) = ...$ và do đó $\alpha = ...$
Như vậy, độ tin cậy là .....

c. Ta có độ chính xác nhỏ hơn 40 giờ với độ tin cậy 98%, tức là

$$z_{\alpha/2}\frac{s}{\sqrt{n}} < 40.$$

Suy ra

$$\sqrt{n} > z_{\alpha/2}\frac{s}{40}$$

Vì $1-\alpha = 0,98$ nên $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2} = 0,99$. Suy ra $z_{\alpha/2} =$ .....Do đó

$$\sqrt{n} > z_{\alpha/2}\frac{s}{40} = 2,33\frac{217,3683}{40} = ...$$

Như vậy, $n >$ ..... và do đó cần khảo sát ít nhất..... bóng đèn.

Bài tập 3. Một thống kê cho thấy chi phí (tính bằng triệu) của các mẫu quảng cáo 30-giây trên một số đài truyền hình được cho như sau

14 55 165 9 15 66 23 30 150 22 12 13 54 73 55 41 78

Giả sử chi phí cho một video quảng cáo 30-giây có phân phối chuẩn. Ước tính chi phí trung bình cho một quảng cáo 30-giây trên truyền hình với độ tin cậy 90%.

Tìm khoảng ước lượng cho trung bình tổng thể với độ tin cậy $1-\alpha$

1. Biết độ lệch chuẩn σ:

- Tính $\overline{x}$,
- Tìm $z_{\alpha/2}$

$\Rightarrow$ khoảng ước lượng cho trung bình tổng thể với độ tin cậy $1-\alpha$ là

$$\boxed{[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]}$$

Tìm khoảng ước lượng cho trung bình tổng thể với độ tin cậy $1-\alpha$

2. Không biết độ lệch chuẩn σ:

- Tính $\overline{x}$ và s;
- Tính sai số của ước lượng:
  1. $n \geq 30$: Tìm $z_{\alpha/2} \Rightarrow$ khoảng ước lượng cho trung bình tổng thể với độ tin cậy $1-\alpha$ là

$$\boxed{[\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}]}$$

  2. $n < 30$: Tìm $t_{\alpha/2}$ bậc $n-1 \Rightarrow$ khoảng ước lượng cho trung bình tổng thể với độ tin cậy $1-\alpha$ là

$$\boxed{[\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}]}$$

5.2.3 Ước lượng tỉ lệ của tổng thể

Bài toán 4. Thăm dò ý kiến của 100 cử tri được chọn ngẫu nhiên tại một thành phố cho thấy có 80% trong số cử tri này ủng hộ ứng viên A. Với độ tin cậy 98%, hãy ước lượng tỉ lệ của tất cả các cử tri ủng hộ ứng viên A tại thành phố này.

- p : tỉ lệ tổng thể (tỉ lệ phần tử có tính chất $\mathcal{P}$ trong tổng thể)
- f : tỉ lệ mẫu cụ thể (tỉ lệ phần tử có tính chất $\mathcal{P}$ trong mẫu)
- Khi $nf > 5$ và $n(1-f) > 5$ thì tỉ lệ mẫu ngẫu nhiên sẽ xấp xỉ phân phối chuẩn $N(f; \dfrac{f(1-f)}{n})$.
- Với độ tin cậy $1-\alpha$, khoảng tin cậy chứa tỉ lệ tổng thể là

$$[f - z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}; f + z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}]$$

  trong đó $\Phi(z_{\alpha/2}) = 1 - \dfrac{\alpha}{2}$ (xem bảng A4).
- Độ chính xác (sai số) là $z_{\alpha/2}\sqrt{\dfrac{f(1-f)}{n}}$.

Theo bất đẳng thức Cauchy, ta có độ chính xác (sai số)

$$z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} \leq z_{\alpha/2}\frac{1}{2\sqrt{n}}.$$

Do đó, sai số tối đa trong ước lượng tỉ lệ tổng thể là $\frac{z_{\alpha/2}}{2\sqrt{n}}$.

Giải Bài toán 4. Theo đề bài

- Tỉ lệ mẫu cụ thể $f = ...$
- Kích thước mẫu $n = ...$
- Độ tin cậy $1 - \alpha = ...$ suy ra $1 - \dfrac{\alpha}{2} = ...$ Do đó $z_{\alpha/2} = ...$
- Độ chính xác (sai số) là

Giải Bài toán 4. Theo đề bài

- Tỉ lệ mẫu cụ thể $f = ...$
- Kích thước mẫu $n = ...$
- Độ tin cậy $1 - \alpha = ...$ suy ra $1 - \dfrac{\alpha}{2} = ...$ Do đó $z_{\alpha/2} = ...$
- Độ chính xác (sai số) là

$$z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} = 2,33\frac{0,4}{10} = 0,0932.$$

- Khoảng tin cậy $[0,7068; 0,8932]$.

Như vậy có từ 70,68% đến 89,32% cử tri ủng hộ ứng viên A.

Bài tập 4. Lấy ngẫu nhiên 200 sản phẩm trong một kho hàng để kiểm tra thì thấy có 21 sản phẩm có lỗi.

a. Với độ tin cậy 95%, hãy ước lượng tỉ lệ sản phẩm lỗi của cả kho hàng.

b. Dựa vào mẫu trên, để ước tính tỉ lệ sản phẩm bị lỗi có độ chính xác là 0,035 thì độ tin cậy bằng bao nhiêu?

c. Dựa vào mẫu trên, nếu muốn ước lượng tỉ lệ sản phẩm bị lỗi với độ chính xác nhỏ hơn 0,01 với độ tin cậy 93% thì cần kiểm tra ít nhất bao nhiêu sản phẩm.

5.3 Kiểm định giả thuyết thống kê

Bài toán

Một hiệu trưởng của một trường THPT tại TPHCM đọc báo thấy rằng điểm trung bình của bài thi Đánh giá năng lực đợt 1 của Đại học Quốc gia TPHCM năm 2022 là 646,1 điểm. Hiệu trưởng nói rằng điểm trung bình của tất cả các học sinh của trường thi đáng giá năng lực lớn hơn 646,1. Sau đó, một phóng viên chọn ngẫu nhiên 50 học sinh của trường đã thi Đánh giá năng lực và thấy rằng điểm trung bình của nhóm học sinh này là 665. Như vậy, có đủ căn cứ để chấp nhận phát biểu của hiệu trưởng không?

Bài toán

Một hiệu trưởng của một trường THPT tại TPHCM đọc báo thấy rằng điểm trung bình của bài thi Đánh giá năng lực đợt 1 của Đại học Quốc gia TPHCM năm 2022 là 646,1 điểm. Hiệu trưởng nói rằng điểm trung bình của tất cả các học sinh của trường thi đáng giá năng lực lớn hơn 646,1. Sau đó, một phóng viên chọn ngẫu nhiên 50 học sinh của trường đã thi Đánh giá năng lực và thấy rằng điểm trung bình của nhóm học sinh này là 665. Như vậy, có đủ căn cứ để chấp nhận phát biểu của hiệu trưởng không?

Định nghĩa 5.3.1 (Giả thuyết thống kê)

Giả thuyết thống kê là một dự đoán về một tham số của tổng thể.

Định nghĩa 5.3.2

1. **Giả thuyết** (null hypothesis), kí hiệu H_0, là một giả thuyết thống kê nói rằng **không có sự khác biệt** giữa một tham số và một giá trị cụ thể hoặc không có sự khác biệt giữa hai tham số.
2. **Đối thuyết** (alternative hypothesis), ký hiệu H_1, là một giả thuyết thống kê cho biết **có sự khác biệt** giữa một tham số và một giá trị cụ thể, hoặc có sự khác biệt giữa hai tham số.

Định nghĩa 5.3.2

1. **Giả thuyết** (null hypothesis), kí hiệu H_0, là một giả thuyết thống kê nói rằng **không có sự khác biệt** giữa một tham số và một giá trị cụ thể hoặc không có sự khác biệt giữa hai tham số.
2. **Đối thuyết** (alternative hypothesis), ký hiệu H_1, là một giả thuyết thống kê cho biết **có sự khác biệt** giữa một tham số và một giá trị cụ thể, hoặc có sự khác biệt giữa hai tham số.

Kết quả của mỗi kiểm định là *chấp nhận* H_0 **hoặc** *bác bỏ* H_0 *và chấp nhận* H_1.

Định nghĩa 5.3.2

1. **Giả thuyết** (null hypothesis), kí hiệu H_0, là một giả thuyết thống kê nói rằng **không có sự khác biệt** giữa một tham số và một giá trị cụ thể hoặc không có sự khác biệt giữa hai tham số.
2. **Đối thuyết** (alternative hypothesis), ký hiệu H_1, là một giả thuyết thống kê cho biết **có sự khác biệt** giữa một tham số và một giá trị cụ thể, hoặc có sự khác biệt giữa hai tham số.

Kết quả của mỗi kiểm định là *chấp nhận* H_0 **hoặc** *bác bỏ* H_0 *và chấp nhận* H_1.

Các dạng toán kiểm định thường gặp:

Kiểm định 2 phía: Giả thuyết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu \neq \mu_0$.

Kiểm định 1 phía trái: Giả thuyết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu < \mu_0$.

Kiểm định 1 phía phải: Giả thuyết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu > \mu_0$.

Ví dụ:

1. Một nhà nghiên cứu nói rằng những trẻ em uống ít nhất 1 ly sữa mỗi ngày khi trưởng thành sẽ có chiều cao lớn hơn 170cm.
   Ta kiểm định: Giả thuyết $H_0 : \mu = 170$ và đối thuyết $H_1 : \mu > 170$.
2. Một giám đốc của một doanh nghiệp thấy rằng sau dịch Covid-19 mức lương trung bình của công nhân toàn công ty có thay đổi. Mức lương trung bình trước dịch Covid-19 là 8,2 triệu đồng/tháng.
   Ta kiểm định: Giả thuyết $H_0 : \mu = 8,2$ và đối thuyết $H_1 : \mu \neq 8,2$.
3. Một công nhân sản xuất gạch thấy rằng số lượng gạch làm ra trong 1 giờ giảm khi áp dụng quy trình sản xuất mới. Trước đây, trung bình công nhân làm được 35 viên gạch trong một giờ.
   Ta kiểm định: Giả thuyết $H_0 : \mu = 35$ và đối thuyết $H_1 : \mu < 35$.
4. Một nhân viên của một nhà hàng nói rằng thời gian trung bình khách phải chờ để được phục vụ của nhà hàng họ là không quá 10 phút.
   Ta kiểm định: Giả thuyết $H_0 : \mu = 10$ và đối thuyết $H_1 : \mu > 10$.

Các sai lầm trong kiểm định giả thuyết

- Sai lầm loại 1: H_0 đúng nhưng bác bỏ H_0
- Sai lầm loại 2: H_0 sai nhưng chấp nhận H_0

Trong thực tế, sai lầm loại 1 là nguy hiểm hơn, do đó ta thiết kế mô hình kiểm định sao cho xác suất sai lầm loại 1 bị chặn bởi một số rất nhỏ α.

Định nghĩa 5.3.3 (Mức ý nghĩa - level of significance)

Số α được gọi là **mức ý nghĩa** của kiểm định nếu α là xác suất ta bác bỏ H_0 khi H_0 đúng.

Giả sử biến ngẫu nhiên X có phân phối chuẩn $N(\mu;\sigma^2)$ trong đó μ là trung bình của tổng thể.

Bài toán 1. Ta kiểm định

$$H_0 : \mu = \mu_0 \text{ và } H_1 : \mu \neq \mu_0.$$

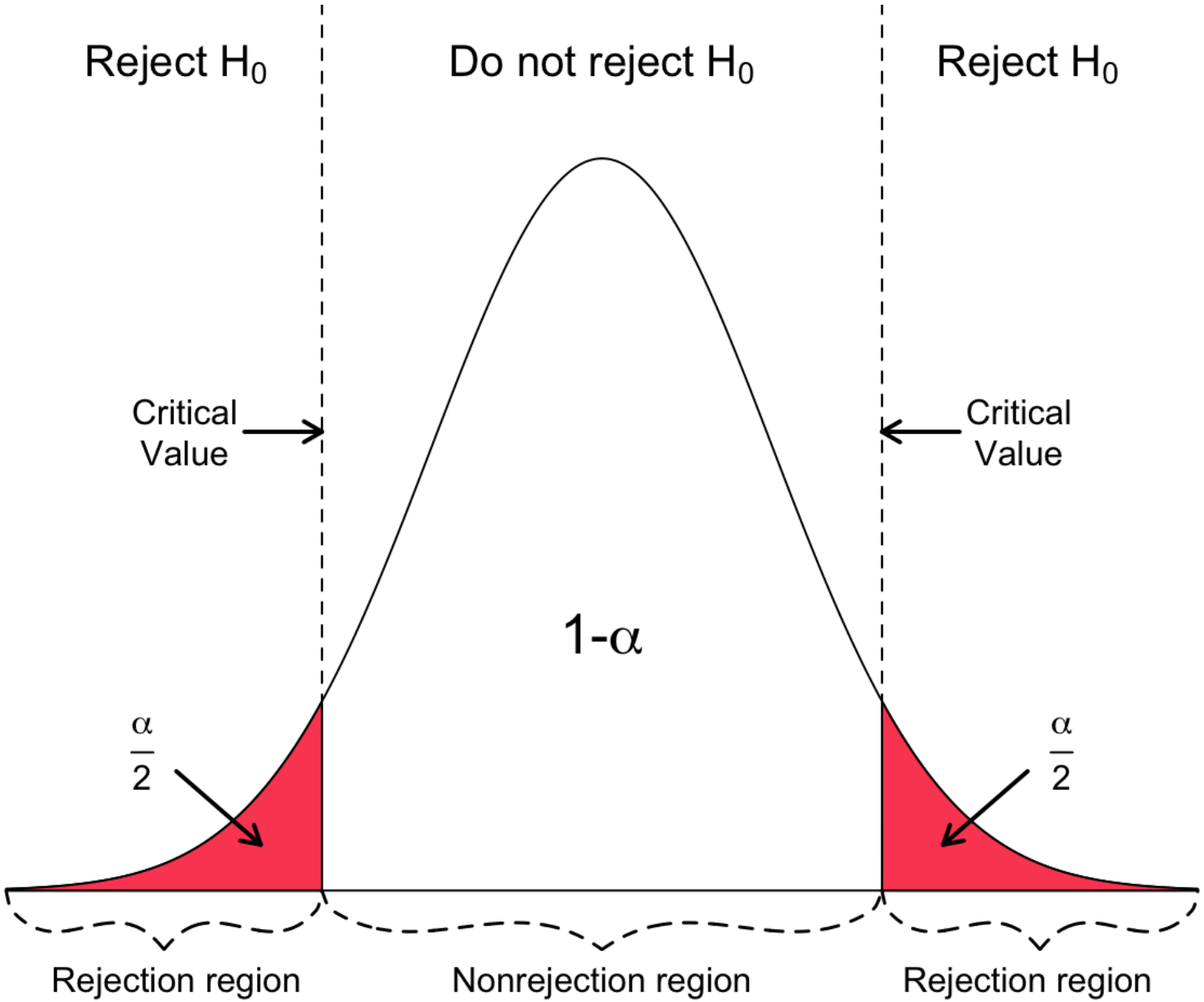
Đặt $Z = \dfrac{\overline{X} - \mu_0}{\sigma}$ là biến ngẫu nhiên có phân phối chuẩn $N(0;1)$.

Với mức ý nghĩa α, đặt $z_{\alpha/2}$ (giá trị tới hạn) là giá trị thỏa mãn

$$P(|Z| > z_{\alpha/2}) = \alpha$$

hay

$$\boxed{\Phi(z_{\alpha/2}) = P(Z \leq z_{\alpha/2}) = 1 - \frac{\alpha}{2}}$$


Reject H_0
Do not reject H_0
Reject H_0
Critical Value
Critical Value
1-α
α/2
α/2
Rejection region
Nonrejection region
Rejection region

Bài toán 2. Với mức ý nghĩa α, ta kiểm định:

$$H_0 : \mu = \mu_0 \text{ và } H_1 : \mu > \mu_0.$$

Đặt z_α là giá trị thỏa mãn

$$P(Z > z_\alpha) = \alpha$$

hay

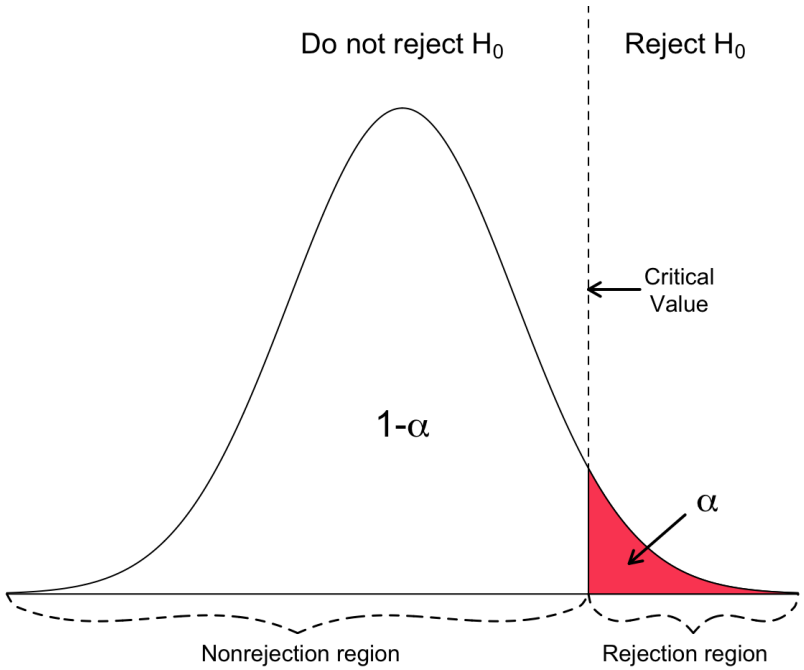
$$\boxed{\Phi(z_\alpha) = P(Z \leq z_\alpha) = 1 - \alpha}$$

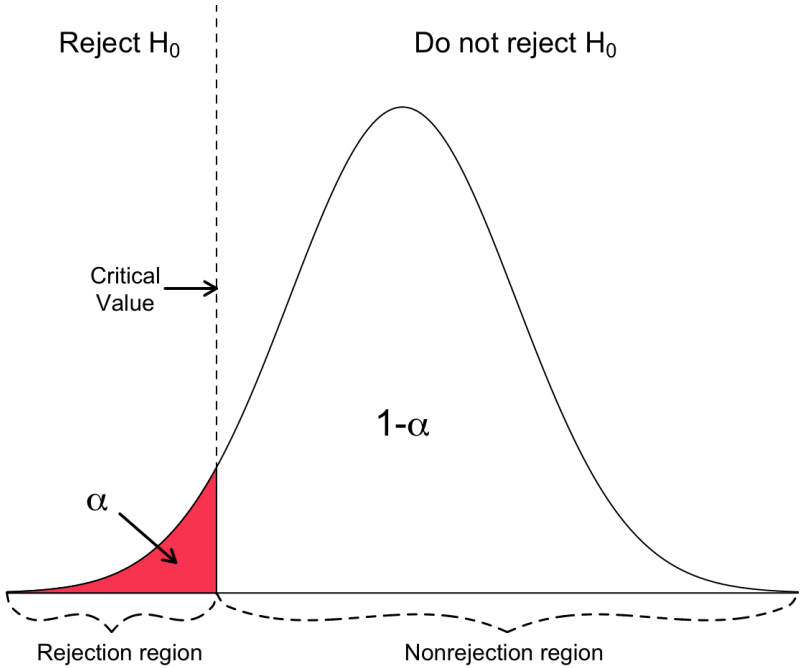
Bài toán 3. Với mức ý nghĩa α, ta kiểm định:

$$H_0 : \mu = \mu_0 \text{ và } H_1 : \mu < \mu_0.$$

Đặt z_α là giá trị thỏa mãn

$$\boxed{\Phi(z_\alpha) = P(Z < z_\alpha) = \alpha}$$


Do not reject H_0
Reject H_0
Critical Value
1-α
α
Nonrejection region
Rejection region


Reject H_0
Do not reject H_0
Critical Value
1-α
α
Rejection region
Nonrejection region

5.3.1 Kiểm định giả thuyết về trung bình

Cho $\overline{x}$ là trung bình mẫu, n là kích thước mẫu, s là độ lệch chuẩn mẫu hiệu chỉnh.

Trường hợp 1. σ đã biết.	**Trường hợp 2.** σ chưa biết và kích thước mẫu $n \geq 30$.
Tính $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	Tính $z = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

5.3.1 Kiểm định giả thuyết về trung bình

Cho $\overline{x}$ là trung bình mẫu, n là kích thước mẫu, s là độ lệch chuẩn mẫu hiệu chỉnh.

Trường hợp 1. σ đã biết.	**Trường hợp 2.** σ chưa biết và kích thước mẫu $n \geq 30$.
Tính $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	Tính $z = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

Kiểm định	Bác bỏ H_0	Chấp nhận H_0
$H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0$	$\lvert z \rvert \geq z_{\alpha/2}$	$\lvert z \rvert < z_{\alpha/2}$
$H_0 : \mu = \mu_0; H_1 : \mu > \mu_0$	$z \geq z_\alpha$, với $\Phi(z_\alpha) = 1 - \alpha$	$z < z_\alpha$
$H_0 : \mu = \mu_0; H_1 : \mu < \mu_0$	$z \leq z_\alpha$, với $\Phi(z_\alpha) = \alpha$	$z > z_\alpha$

Trường hợp 3: Với σ chưa biết và $n < 30$, tổng thể có phân phối chuẩn. Đặt

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim St(n)$$

trong đó $\overline{X}$ là trung bình mẫu ngẫu nhiên, n là cỡ mẫu và s là độ lệch chuẩn mẫu hiệu chỉnh.

Trường hợp 3: Với σ chưa biết và $n < 30$, tổng thể có phân phối chuẩn. Đặt

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim St(n)$$

trong đó $\overline{X}$ là trung bình mẫu ngẫu nhiên, n là cỡ mẫu và s là độ lệch chuẩn mẫu hiệu chỉnh.

Với mức ý nghĩa α, đặt $t, t_{\alpha/2}$ và t_α là các số thực thỏa mãn (xem bảng A5)

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}, \quad P(T > t_\alpha) = \alpha, \quad P(T > t_{\alpha/2}) = \frac{\alpha}{2}.$$

Trường hợp 3: Với σ chưa biết và $n < 30$, tổng thể có phân phối chuẩn. Đặt

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim St(n)$$

trong đó $\overline{X}$ là trung bình mẫu ngẫu nhiên, n là cỡ mẫu và s là độ lệch chuẩn mẫu hiệu chỉnh.

Với mức ý nghĩa α, đặt $t, t_{\alpha/2}$ và t_α là các số thực thỏa mãn (xem bảng A5)

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}, \quad P(T > t_\alpha) = \alpha, \quad P(T > t_{\alpha/2}) = \frac{\alpha}{2}.$$

Kiểm định	Bác bỏ H_0	Chấp nhận H_0
$H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0.$	$\lvert t\rvert \geq t_{\alpha/2}$	$\lvert t\rvert < t_{\alpha/2}$
$H_0 : \mu = \mu_0; H_1 : \mu > \mu_0.$	$t \geq t_\alpha$	$t < t_\alpha$
$H_0 : \mu = \mu_0; H_1 : \mu < \mu_0.$	$t \leq -t_\alpha$	$t > -t_\alpha$

Ví dụ 5.3.3

Theo báo cáo "Thị trường IT Việt Nam - Developers Recruitment State 2021" do TopDev công bố cho biết, tính đến quý II/2021, kỹ sư trí tuệ nhân tạo (AI) và máy học (Machine Learning) là vị trí có mức lương trung bình hàng tháng cao nhất trong các kỹ sư IT, đạt 3054 USD (khoảng 70 triệu đồng). Một cuộc khảo sát 30 kỹ sư trí tuệ nhân tạo tốt nghiệp từ một trường đại học X cho thấy họ có mức lương trung bình là 3105 USD/tháng. Hãy kiểm tra kết luận nói rằng các kỹ sư trí tuệ nhân tạo của trường X có mức thu nhập trung bình lớn hơn 3054 USD/tháng với mức ý nghĩa 0,05. Giả sử thu nhập của các kỹ sư trí tuệ nhân tạo có phân phối chuẩn với độ lệch chuẩn tổng thể là 120 USD.

Giải.

- Gọi μ thu nhập trung bình của các kỹ sư trí tuệ nhân tạo
- Ta kiểm định: Giả thuyết $H_0 : \mu = 3054$ và đối thuyết $H_1 : \mu > 3054$
- Theo đề bài, trung bình mẫu là $\overline{x} = 3105$, cỡ mẫu $n = 30$ và độ lệch chuẩn tổng thể $\sigma = 120$
- Vì mức ý nghĩa $\alpha = 0,05$ nên $z_\alpha = 1,65$.
- Đặt

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{3105 - 3054}{120/\sqrt{30}} = 2,34.$$

- Vì $z = 2,34 > 1,65$ nên bác bỏ H_0.
- Ta có căn cứ để đồng ý với tuyên bố lương trung bình của các kỹ sư trí tuệ nhân tạo nhiều hơn 3054 USD/tháng.

Ví dụ 5.3.4

Một nhà nghiên cứu nói rằng trung bình giá tiền của một đôi giày thể thao nam là ít hơn 80 USD. Chọn ngẫu nhiên 36 đôi giày thể thao nam để khảo sát giá, ta được kết quả sau (USD/đôi)

60	70	75	55	80	55	50	40	80
70	50	95	120	90	75	85	80	60
110	65	80	85	85	45	75	60	90
90	60	95	110	85	45	90	70	70

Giả sử giá giày có phân phối chuẩn với độ lệch chuẩn là 19,2 USD. Tuyên bố của nhà nghiên cứu có chấp nhận được không với mức ý nghĩa 10%?

Giải.

- Gọi μ giá trung bình của một đôi giày thể thao nam.
- Ta điểm định: Giả thuyết $H_0 : \mu = 80$ và đối thuyết $H_1 : \mu < 80$
- Theo đề bài, trung bình mẫu là $\overline{x} = 75$, cỡ mẫu $n = 36$ và độ lệch chuẩn tổng thể $\sigma = 19,2$
- Vì mức ý nghĩa $\alpha = 0,1$ nên $z_\alpha = -1,28$.
- Đặt

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{75 - 80}{19,2/\sqrt{36}} = -1,56.$$

- Vì $z = -1,56 < -1,28$ nên bác bỏ H_0.
- Ta có căn cứ để đồng ý với nhận xét giá tiền trung bình của một đôi giày thể thao nam ít hơn 80 USD.

Ví dụ 5.3.5

Một nhà nghiên cứu nói rằng trung bình giá tiền của một đôi giày thể thao nam là ít hơn 80 USD. Chọn ngẫu nhiên 16 đôi giày thể thao nam để khảo sát giá, ta được kết quả sau (USD/đôi)

60	70	75	55	80	55	50	40
70	50	95	120	90	75	85	80

Giả sử giá giày có phân phối chuẩn. Tuyên bố của nhà nghiên cứu có chấp nhận được không với mức ý nghĩa 10%?

Giải.

- Gọi μ giá trung bình của một đôi giày thể thao nam.
- Ta điểm định: Giả thuyết $H_0 : \mu = 80$ và đối thuyết $H_1 : \mu < 80$
- Theo đề bài, trung bình mẫu là $\overline{x} = 71,875$, cỡ mẫu $n = 16$ và độ lệch chuẩn mẫu hiệu chỉnh $s =$ .....
- Vì mức ý nghĩa $\alpha = 0,1$ nên $t_\alpha =$ ..... (bậc tự do 15).
- Đặt

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{.....}{...../\sqrt{.....}} = ...$$

- Vì ..... nên .....
- ....

5.3.2 Kiểm định giả thuyết về tỉ lệ

Bài toán 4

Một người ăn kiêng nói rằng có 60% số người không ăn bánh ngọt. Một cuộc khảo sát 200 người, ta thấy có 128 người nói rằng họ không ăn bánh ngọt. Với mức ý nghĩa 5%, ta có thể bác bỏ tuyên bố của người ăn kiêng này không?

- Gọi p (f, F) là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong tổng thể (trong mẫu cụ thể, mẫu ngẫu nhiên).
- Kiểm định giả thuyết

$$H_0 : p = p_0$$

- Chọn một mẫu ngẫu nhiên có kích thước n.
- Với n đủ lớn, biến ngẫu nhiên

$$Z = \frac{F - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0;1)$$

Đặt

$$z = \frac{f - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Với mức ý nghĩa α.

Kiểm định	Bác bỏ H_0	Chấp nhận H_0
$H_0 : p = p_0; H_1 : p \neq p_0$	$\lvert z\rvert \geq z_{\alpha/2}$	$\lvert z\rvert < z_{\alpha/2}$
$H_0 : p = p_0; H_1 : p > p_0$	$z \geq z_\alpha$, với $\Phi(z_\alpha) = 1 - \alpha$	$z < z_\alpha$
$H_0 : p = p_0; H_1 : p < p_0$	$z \leq z_\alpha$, với $\Phi(z_\alpha) = \alpha$	$z > z_\alpha$

Bài toán 4: Cho $n = 200; p_0 = 60\%$ và mức ý nghĩa 5%.

Giải Bài toán 4

- Gọi p là tỉ lệ người không ăn bánh ngọt.
- Ta kiểm định: Giả thuyết $H_0 : p = 60\%$ và đối thuyết $H_1 : p \neq 60\%$
- Tỉ lệ mẫu là $f = \dfrac{128}{200} = 0,64$ và cỡ mẫu $n = 200$.
- Vì mức ý nghĩa $\alpha = 0,05$ nên $z_{\alpha/2} = 1,96$.
- Đặt

$$z = \frac{f - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0,64 - 0,6}{\sqrt{0,6(1-0,6)/200}} = 1,15.$$

- Vì $|z| = 1,15 < 1,96$ nên chấp nhận H_0.
- Ta đồng ý với phát biểu rằng có 60% người không ăn bánh ngọt.

Bài tập 2

Một giáo viên nói rằng lương trung bình của giáo viên tại TPHCM ít hơn 6 triệu đồng/tháng trong năm 2021. Chọn ngẫu nhiên 8 giáo viên thì thấy mức lương hàng tháng (đơn vị là triệu đồng) của họ trong năm 2021 là

$$6 \quad 5{,}6 \quad 6 \quad 5{,}5 \quad 7 \quad 5{,}5 \quad 6 \quad 5{,}5$$

Giả sử lương của giáo viên có phân phối chuẩn. Với mức ý nghĩa 10%, tuyên bố của giáo viên đó có chấp nhận được không?

Bài tập 3

Một lập trình viên nói rằng có hơn 25% các lập trình viên đã học ngôn ngữ lập trình Python. Một cuộc khảo sát 200 lập trình viên tại một thành phố nọ, người ta thấy có 63 lập trình viên đã học Python. Với mức ý nghĩa 5%, hãy kết luận về nhận định của lập trình viên trên.

Giải Bài tập 2.

- Gọi μ tiền lương trung bình hàng tháng của giáo viên trong năm 2021.
- Ta kiểm định: Giả thuyết $H_0 : \mu = ...$ và đối thuyết $H_1 : \mu .....$
- Trung bình mẫu là $\overline{x} = ...$ cỡ mẫu $n = 8$ và độ lệch chuẩn mẫu hiệu chỉnh là $s = ...$
- Mức ý nghĩa $\alpha = 0,1$ suy ra ..... (xem bảng A5)

Giải Bài tập 2.

- Gọi μ tiền lương trung bình hàng tháng của giáo viên trong năm 2021.
- Ta kiểm định: Giả thuyết $H_0 : \mu = ...$ và đối thuyết $H_1 : \mu.....$
- Trung bình mẫu là $\overline{x} = ...$ cỡ mẫu $n = 8$ và độ lệch chuẩn mẫu hiệu chỉnh là $s = ...$
- Mức ý nghĩa $\alpha = 0,1$ suy ra ..... (xem bảng A5)
- Đặt

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{5,888 - 6}{0,508/\sqrt{8}} = -0,624.$$

- Vì $t = -0,624 > -1,415$ nên chấp nhận H_0.
- Ta không đồng ý với tuyên bố tiền lương trung bình trong một tháng của giáo viên ít hơn 6 triệu đồng.

5.3.3 So sánh hai trung bình, hai tỉ lệ

Bài toán

Một cuộc khảo sát trong năm 2021 ở Việt Nam cho thấy rằng mức lương trung bình của 50 kỹ sư phần mềm là 1840 USD/tháng và 50 quản lý IT là 1750 USD/tháng. Giả sử độ lệch chuẩn tổng thể lần lượt là 400 USD và 370 USD triệu. Với mức ý nghĩa 5%, ta có thể kết luận rằng có sự khác biệt giữa mức lương trung bình của hai nhóm công việc trong ngành công nghệ thông tin không?

5.3.3 So sánh hai trung bình, hai tỉ lệ

Bài toán

Một cuộc khảo sát trong năm 2021 ở Việt Nam cho thấy rằng mức lương trung bình của 50 kỹ sư phần mềm là 1840 USD/tháng và 50 quản lý IT là 1750 USD/tháng. Giả sử độ lệch chuẩn tổng thể lần lượt là 400 USD và 370 USD triệu. Với mức ý nghĩa 5%, ta có thể kết luận rằng có sự khác biệt giữa mức lương trung bình của hai nhóm công việc trong ngành công nghệ thông tin không?

Đây là dạng toán so sánh hai trung bình. Ta có giả thuyết và đối thuyết lần lượt là

$$H_0 : \mu_X = \mu_Y;\ H_1 : \mu_X \neq \mu_Y.$$

trong đó μ_X, μ_Y lần là thu nhập trung bình của kỹ sư phần mềm và quản lý IT.

Giả thuyết và đối thuyết:

$$H_0 : \mu_X - \mu_Y = 0;\ H_1 : \mu_X - \mu_Y \neq 0.$$

Bài toán so sánh hai trung bình của tổng thể cần các điều kiện:

- Các mẫu của hai tổng thể phải độc lập.
- Biết độ lệch chuẩn của các tổng thể
- Nếu các kích thước mẫu nhỏ hơn 30 thì các tổng thể phải có phân phối chuẩn hoặc xấp xỉ phân phối chuẩn.

Đặt $\overline{X}, \overline{Y}$ lần lượt là các trung bình của hai mẫu ngẫu nhiên được lấy từ hai tổng thể. Độ lệch chuẩn tổng thể lần lượt là σ_X, σ_Y và các kích thước mẫu lần lượt là n_X và n_Y. Khi đó

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}} \sim N(0;1).$$

Từ một bài toán so sánh trung bình hai tổng thể, ta chuyển về bài toán kiểm định trung bình của tổng thể.

Các dạng toán so sánh trung bình như sau:

So sánh	Chuyển thành
$H_0 : \mu_X = \mu_Y;\ H_1 : \mu_X \neq \mu_Y$	$H_0 : \mu_X - \mu_Y = 0;\ H_1 : \mu_X - \mu_Y \neq 0$
$H_0 : \mu_X = \mu_Y;\ H_1 : \mu_X < \mu_Y$	$H_0 : \mu_X - \mu_Y = 0;\ H_1 : \mu_X - \mu_Y < 0$
$H_0 : \mu_X = \mu_Y;\ H_1 : \mu_X > \mu_Y$	$H_0 : \mu_X - \mu_Y = 0;\ H_1 : \mu_X - \mu_Y > 0$

5.4 Tương quan và hồi quy tuyến tính

Trong Machine Learning, một trong những thuật toán quan trọng nhất là Thuật toán Hồi quy tuyến tính (Linear Regression) thuộc nhóm *Học có giám sát* (Supervised Learning).

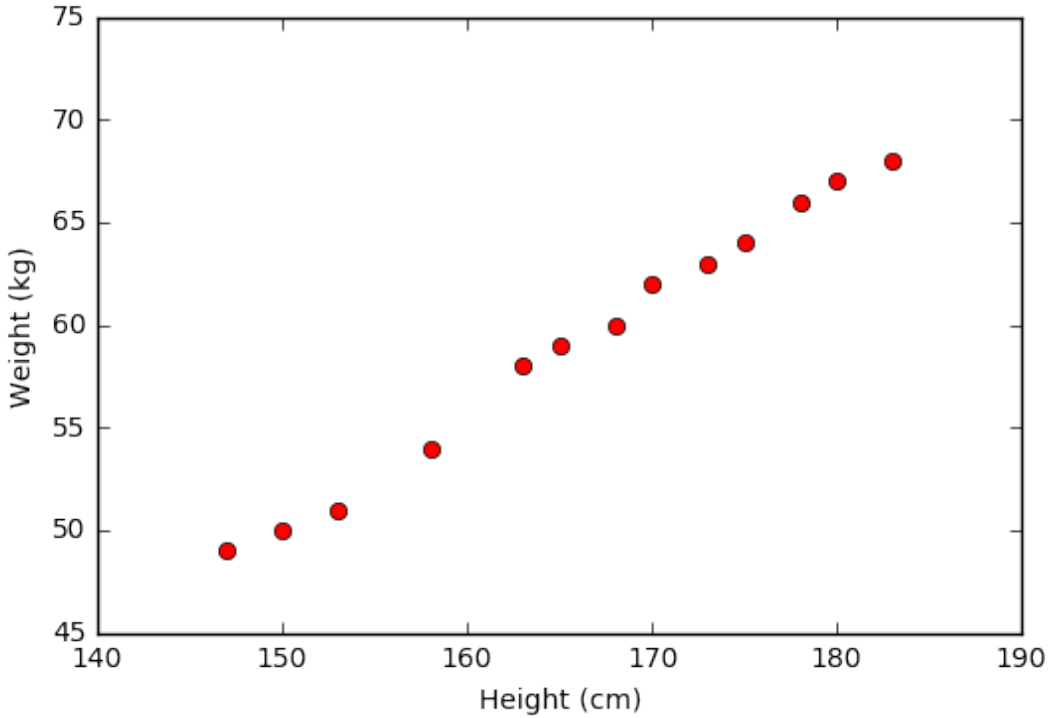
Bài toán

Bảng dữ liệu về chiều cao và cân nặng của 15 người:

Chiều cao (cm)	Cân nặng (kg)	Chiều cao (cm)	Cân nặng (kg)
147	49	168	60
150	50	170	72
153	51	173	63
155	52	175	64
158	54	178	66
160	56	180	67
163	58	183	68
165	59		

Có thể dự đoán cân nặng của một người dựa vào chiều cao của họ không?

Biểu diễn các dữ liệu trên dưới dạng đồ thị như sau

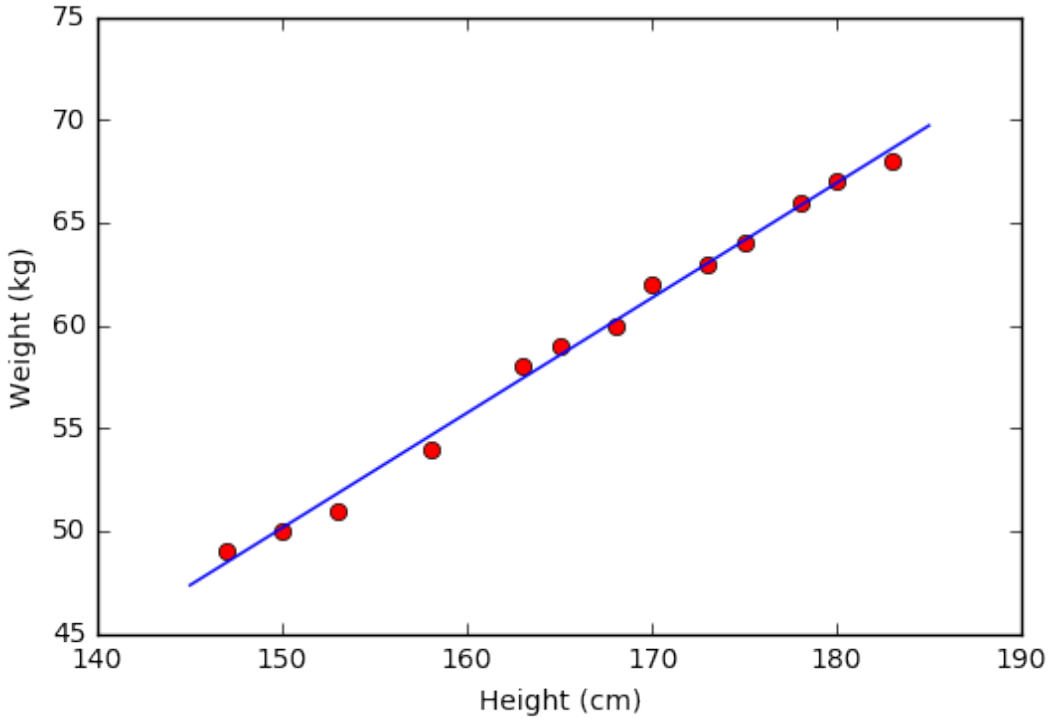


Ta thấy rằng dữ liệu được sắp xếp gần như theo một đường thẳng. Do đó mô hình Hồi quy tuyến tính (Linear Regression) nhiều khả năng sẽ cho kết quả tốt. Ta có thể đưa ra mối liên hệ giữa cân nặng và chiều cao như sau

$$\text{cân nặng} = B \times \text{chiều cao} + A.$$

Bằng các công cụ tính toán, chúng ta sẽ tính được A, B.

Khi đó, các điểm dữ liệu nằm khá gần đường thẳng mà ta dự đoán.

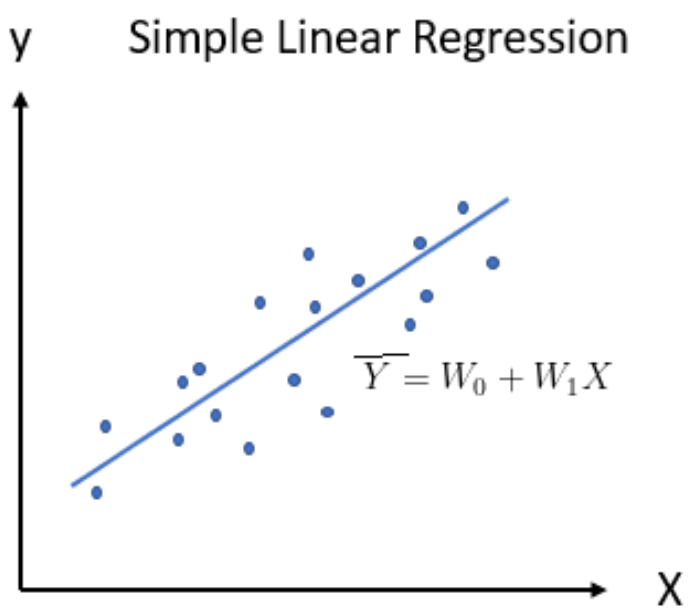


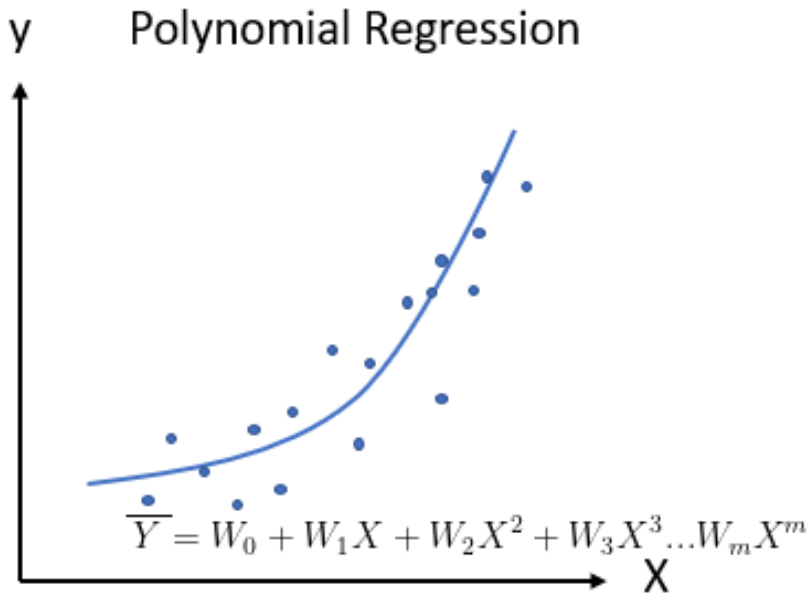
Sử dụng mô hình này, ta có thể dự đoán cân nặng của một người có chiều cao 155cm, 160 cm hoặc 171cm.

5.4.1 Hệ số tương quan

Xét vectơ ngẫu nhiên (X, Y) và tập n giá trị cụ thể $(x_1, y_1), \ldots, (x_n, y_n)$. Các cặp giá trị này được gọi là **dữ liệu thực nghiệm**.
Tập hợp các điểm (x_i, y_i) được biểu diễn trên mặt phẳng tọa độ được gọi là **biểu đồ phân tán** (Scatter diagram).
Có nhiều kiểu phụ thuộc giữa hai biến ngẫu nhiên X và Y nhưng phổ biến nhất là dạng phụ thuộc hàm số $Y = f(X)$. Một trong những hàm đơn giản nhất là hàm số bậc nhất $Y = aX + b$ hay dạng tuyến tính.
Đường cong phù hợp là một đường cong xấp xỉ tốt nhất (ít sai lệch nhất) với các điểm dữ liệu đã cho.

- Nếu đường cong phù hợp là một đường thẳng thì ta có một **quan hệ tuyến tính** (linear relation) giữa hai biến ngẫu nhiên.
- Nếu đường cong phù hợp **không** là một đường thẳng thì ta có một **quan hệ phi tuyến tính** giữa hai biến ngẫu nhiên.





Bài toán

1. Có một quan hệ tuyến tính hoặc phi tuyến tính giữa hai biến ngẫu nhiên không?
2. Nếu có một quan hệ tuyến tính (phi tuyến tính) giữa hai biến ngẫu nhiên thì có thể biểu diễn mối quan hệ này dưới dạng một hàm số không?

Ta cần một số đo để đo mức độ chặt chẽ trong quan hệ tuyến tính giữa hai biến ngẫu nhiên.

Định nghĩa 5.4.1 (Hệ số tương quan - Correlation coefficient)

Hệ số tương quan mẫu của hai biến ngẫu nhiên X, Y

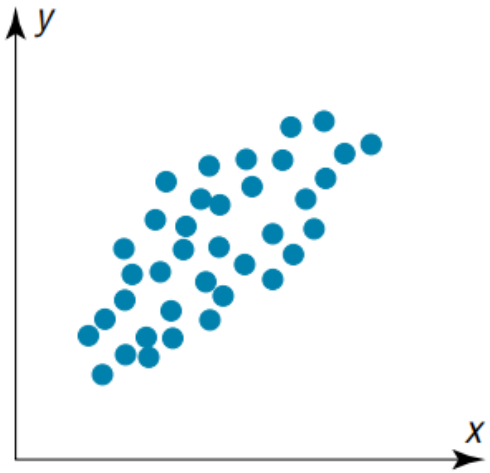
$$r = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2)(\overline{y^2} - \overline{y}^2)}}$$

Hay

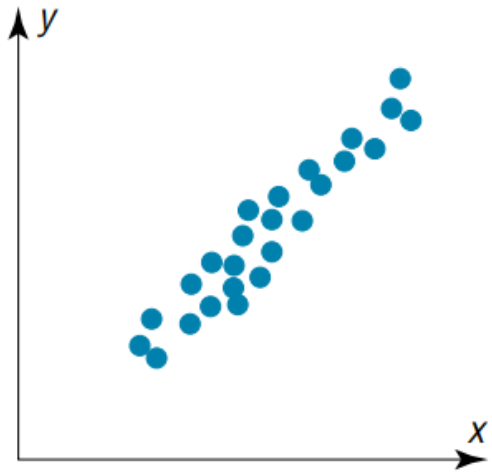
$$r = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i).(\sum_{i=1}^{n} y_i)}{\sqrt{(n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2)(n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2)}}$$

trong đó n là số cặp điểm dữ liệu thực nghiệm.

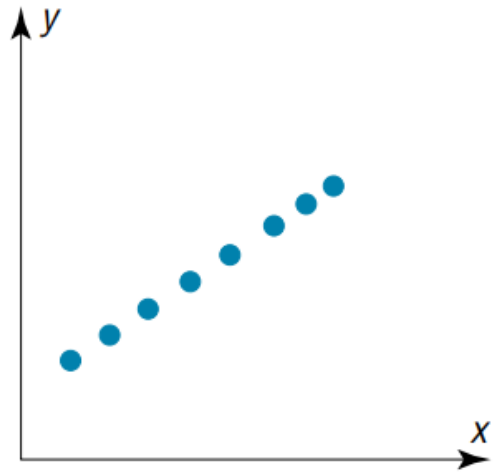
- Ta có $-1 \leq r \leq 1$.
- Hệ số tương quan là một con số đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên.
- Nếu $0,8 \leq |r| \leq 1$ thì ta nói X, Y có tương quan tuyến tính mạnh.
- Nếu $|r| < 0,8$ thì ta nói X, Y có tương quan tuyến tính yếu.
- Nếu r gần bằng 1 thì ta nói có sự tương quan tuyến tính thuận giữa X và Y.
- Nếu r gần bằng -1 thì ta nói có sự tương quan tuyến tính nghịch giữa X và Y.



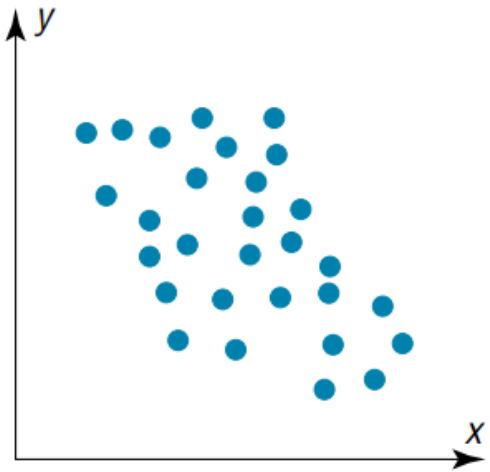
(a) $r = 0.50$



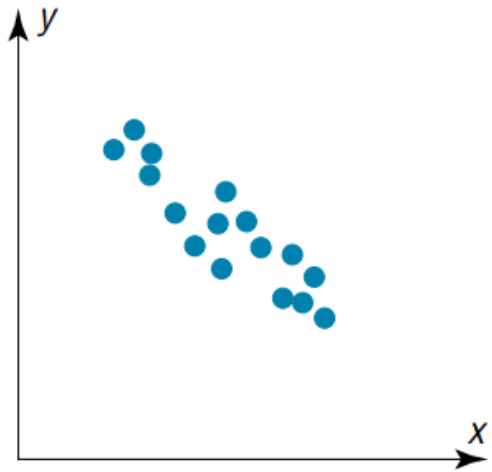
(b) $r = 0.90$



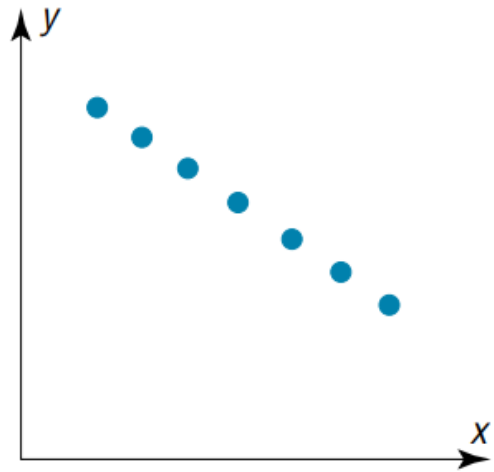
(c) $r = 1.00$



(d) $r = -0.50$



(e) $r = -0.90$



(f) $r = -1.00$

Ví dụ 5.4.2 Điểm số môn Đại số tuyến tính và số buổi vắng của 7 sinh viên được cho bên dưới

Số buổi vắng (X)	6	2	15	9	12	5	8
Điểm (Y)	8,2	8,6	4,3	7,4	5,8	9,0	7,8

Tìm hệ số tương quan giữa số buổi nghỉ học và điểm môn Đại số tuyến tính.

Ví dụ 5.4.2 Điểm số môn Đại số tuyến tính và số buổi vắng của 7 sinh viên được cho bên dưới

Số buổi vắng (X)	6	2	15	9	12	5	8
Điểm (Y)	8,2	8,6	4,3	7,4	5,8	9,0	7,8

Tìm hệ số tương quan giữa số buổi nghỉ học và điểm môn Đại số tuyến tính.

Ta có

$$\overline{xy} = \frac{6.8,2 + 2.8,6 + 15.4,3 + 9.7,4 + 12.5,8 + 5.9,0 + 8.7,8}{7} = 53,5$$

$$\overline{x} = \frac{6 + 2 + 15 + 9 + 12 + 5 + 8}{7} = 8,14$$

$$\overline{y} = \frac{8,2 + 8,6 + 4,3 + 7,4 + 5,8 + 9,0 + 7,8}{7} = 7,3$$

$$\overline{x^2} = \frac{6^2 + 2^2 + 15^2 + 9^2 + 12^2 + 5^2 + 8^2}{7} = 82,71$$

$$\overline{y^2} = \frac{8,2^2 + 8,6^2 + 4,3^2 + 7,4^2 + 5,8^2 + 9,0^2 + 7,8^2}{7} = 55,7$$

Do đó, hệ số tương quan là

$$r = \frac{53,5 - 8,14.7,3}{\sqrt{(82,71 - 8,14^2)(55,7 - 7,3^2)}} = \frac{-5,992}{6,296} = -0,9517.$$

Có một sự tương quan tuyến tính mạnh giữa số buổi vắng và số điểm. Nếu số buổi vắng càng nhiều thì số điểm càng thấp.

5.4.2 Hồi quy

Bài toán

Ta muốn khảo sát xem số buổi nghỉ học có ảnh hưởng đến điểm thi cuối kỳ của môn xác suất thống kê. Nếu biết số buổi nghỉ học thì ta có thể dự đoán điểm thi cuối kỳ được không?

- Mục đích của hồi quy là dự đoán một đại lượng này từ các đại lượng khác.
- Nếu biến Y được ước lượng từ biến X bằng một biểu thức $Y = f(X)$ thì biểu thức này được gọi là **phương trình hồi quy** của Y theo X.
- Đường cong biểu diễn đường $Y = f(X)$ được gọi là **đường cong hồi quy** của Y theo X.
- Đường thẳng biểu diễn đường $Y = A + BX$ (phương trình hồi quy tuyến tính) được gọi là **đường thẳng hồi quy** của Y theo X.

Trong việc nghiên cứu mối liên hệ giữa hai biến:

1. Thu thập dữ liệu và xây dựng biểu đồ phân tán
2. Tính hệ số tương quan r
3. Kiểm tra sự tương quan tuyến tính giữa hai biến
4. Nếu $|r|$ gần bằng 1 thì ta sẽ xác định đường thẳng hồi quy (regression line) (đường thẳng phù hợp nhất).
5. Đường thẳng hồi quy giúp các nhà nghiên cứu có thể nhìn thấy xu hướng và đưa ra các dự báo.

Trong việc nghiên cứu mối liên hệ giữa hai biến:

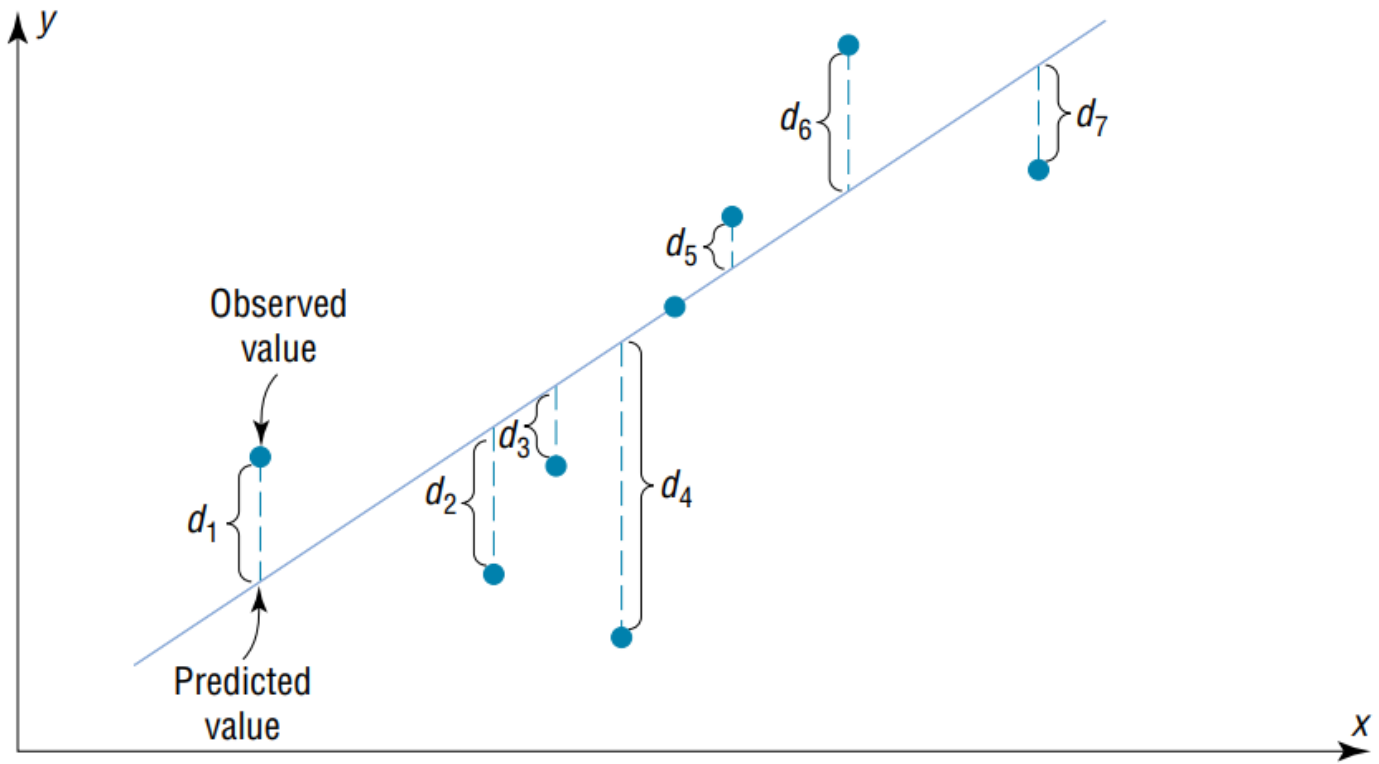
1. Thu thập dữ liệu và xây dựng biểu đồ phân tán
2. Tính hệ số tương quan r
3. Kiểm tra sự tương quan tuyến tính giữa hai biến
4. Nếu $|r|$ gần bằng 1 thì ta sẽ xác định đường thẳng hồi quy (regression line) (đường thẳng phù hợp nhất).
5. Đường thẳng hồi quy giúp các nhà nghiên cứu có thể nhìn thấy xu hướng và đưa ra các dự báo.

Cho các điểm $(x_1, y_1), \ldots, (x_n, y_n)$, ta sẽ tìm phương trình đường thẳng (**hồi quy tuyến tính**) $Y = A + BX$, sao cho

$$\sum_{i=1}^{n} (y_i - (A + Bx_i))^2$$

là nhỏ nhất.

Phương pháp trên được gọi là phương pháp *bình phương cực tiểu* (method of least squares).



Khi đó

$$B = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - \overline{x}^2} \text{ và } A = \overline{y} - B\overline{x}.$$

Ví dụ 5.4.3

Điểm số môn Đại số tuyến tính và số buổi vắng của 7 sinh viên được cho bên dưới

Số buổi vắng (X)	6	2	15	9	12	5	8
Điểm (Y)	8,2	8,6	4,3	7,4	5,8	9,0	7,8

Tìm phương trình đường thẳng hồi quy tuyến tính và dự đoán điểm của sinh viên chỉ vắng 1 buổi học.

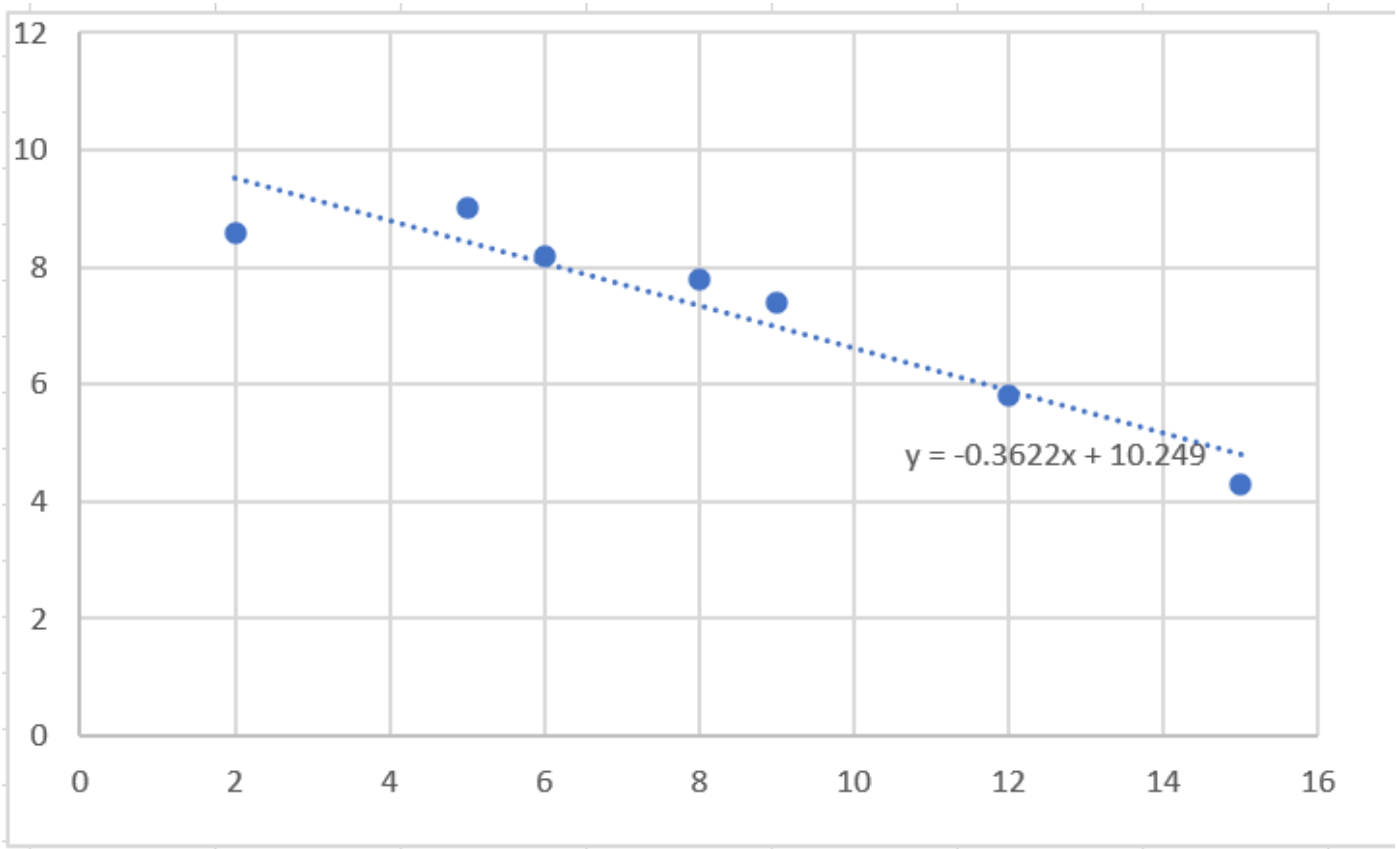
Giải.

- Phương trình hồi quy tuyến tính $Y = A + BX$.
- $B = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - \overline{x}^2} =$ .....
- $A = \overline{y} - B\overline{x} =$ .....
- Phương trình đường thẳng hồi quy tuyến tính cần tìm là .....
- Khi $X = 1$ thì $Y =$ ...

Giải. Dùng máy tính CASIO fx-570VN-PLUS

- $\boxed{SHIFT} \to \boxed{MODE} \to \boxed{\bigtriangledown}$ → chọn STAT (trên màn hình - phím 4)
- Màn hình xuất hiện **Frequency**, chọn **OFF**
- $\boxed{SHIFT} \to \boxed{MODE} \to$ **Data** (phím $\boxed{2}$)
- Nhập dữ liệu cột X : $\boxed{6}\ \boxed{=}\ \boxed{2}\ \boxed{=}$...
- Nhập dữ liệu cột Y : $\boxed{8.2}\ \boxed{=}\ \boxed{8.6}\ \boxed{=}$...
- $\boxed{ON}$
- $\boxed{SHIFT} \to \boxed{1} \to$ **Reg** (phím $\boxed{5}$)
- Chọn **A** (phím $\boxed{1}$) $\boxed{=}$
- $\boxed{ON}$
- $\boxed{SHIFT} \to \boxed{1} \to$ **Reg** (phím $\boxed{5}$)
- Chọn **B** (phím $\boxed{2}$) $\boxed{=}$

Khi đó $A = 10,2493$ và $B = -0,3722$. Đường thẳng hồi quy tuyến tính là $Y = 10,2493 - 0,3622X$.



Nếu $X = 1$ thì $Y = 9,8871$. Do đó nếu sinh viên vắng một buổi học thì điểm số của sinh viên có thể đạt được là 9,8871 điểm.

Dùng Microsoft Excel để tìm đường thẳng hồi quy:

- Tạo bảng dữ liệu trong Microsoft Excel
- Tạo biểu đồ phân tán: Chọn bảng dữ liệu → **Insert** → **Charts** → **All Charts** → **X Y (Scatter)** → **OK**
- Tạo đường thẳng hồi quy: Nhấp vào ⊞ bên góc phải của Chart vừa hiện ra → **Chart Elements**, chọn **Trendline**
- Hiện phương trình đường thẳng hồi quy: Bên cạnh **Trendline** → ▶ **More Options**
- Trong bảng **Format Trendline**, chọn  , kéo xuống bên dưới và chọn **Display Equation on chart**.

Một vài lưu ý

- Đường thẳng hồi quy tuyến tính theo phương pháp bình phương tối tiểu luôn đi qua điểm $(\overline{x}, \overline{y})$
- Khi tính toán cần xác định rõ biến độc lập và biến phụ thuộc
  - Phương tình hồi quy tuyến tính của Y theo X

$$Y = A + BX$$

  - Phương tình hồi quy tuyến tính của X theo Y

$$X = A + BY$$

Bài tập. Bảng khảo sát doanh thu bán hàng online Y và chi phí quảng cáo online X (trong thời gian 15 phút) của 7 cửa hàng được cho như sau: Đơn vị tính là đô la

Doanh số bán hàng	368	340	665	954	331	556	376
Chi phí quảng cáo	1,7	1,5	2,8	5	1,3	2,2	1,3

a. Tính hệ số tương quan và nhận xét về tính tuyến tính của X và Y (mạnh hay yếu).
b. Viết phương trình hồi quy tuyến tính của Y theo X. Dự đoán doanh số bán hàng (trong 15 phút) khi chi phí quảng cáo online trong 15 phút là 4 đô la.