TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC

SUY LUẬN THỐNG KÊ

BÁO CÁO MÔN HỌC

Ngành: HỆ THỐNG THÔNG TIN QUẢN LÝ

Giảng viên: PGS.TS. Nguyễn Thị Thu Thủy

Sinh viên thực hiện: Đặng Minh Anh

Lớp: Hệ thống thông tin quản lý – K64

HÀ NỘI – 2022

NHẬN XÉT CỦA GIẢNG VIÊN

1. Mục tiêu

(a)

(b)

(c)

2. Nội dung

(a)

(b)

(c)

3. Đánh giá kết quả đạt được

(a)

(b)

(c)

Hà Nội, ngày 06 tháng 02 năm 2022

Giảng viên

PGS.TS. Nguyễn Thị Thu Thủy

Lời cảm ơn

Em xin gửi lời cảm ơn đến cô Nguyễn Thị Thu Thuỷ đã hướng dẫn em bộ môn này. Cô đã giúp em có hình dung cụ thể hơn về ngành Phân tích dữ liệu mà em đang quan tâm ạ.

Hà Nội, tháng 02 năm 2022

Sinh viên

Đặng Minh Anh

Tóm tắt nội dung Báo cáo

1.

2.

3. . . .

Mục lục

Mở đầu

Sự cần thiết của học phần

Phân tích dữ liệu rất quan trọng trong nghiên cứu vì nó làm cho việc nghiên cứu dữ liệu trở nên đơn giản và chính xác hơn rất nhiều. Nó giúp các nhà nghiên cứu diễn giải dữ liệu một cách đơn giản để các nhà nghiên cứu không bỏ sót bất kỳ điều gì có thể giúp họ thu thập thông tin chi tiết từ đó.

Phân tích dữ liệu là một cách để nghiên cứu và phân tích lượng dữ liệu khổng lồ. Nghiên cứu thường bao gồm việc duyệt qua hàng đống dữ liệu, ngày càng có nhiều thứ hơn để các nhà nghiên cứu xử lý trong mỗi phút trôi qua.

Do đó, kiến thức phân tích dữ liệu là một lợi thế lớn đối với các nhà nghiên cứu trong thời đại hiện nay, giúp họ hoạt động rất hiệu quả và năng suất.

Phân tích dữ liệu là quá trình phân tích dữ liệu dưới nhiều định dạng khác nhau. Mặc dù ngày nay dữ liệu rất dồi dào, nhưng dữ liệu có sẵn ở các dạng khác nhau và nằm rải rác trên nhiều nguồn khác nhau. Phân tích dữ liệu giúp làm sạch và chuyển đổi tất cả dữ liệu này thành một dạng nhất quán để có thể nghiên cứu hiệu quả.

Một khi dữ liệu được làm sạch, chuyển đổi và sẵn sàng sử dụng, nó có thể làm nên điều kỳ diệu. Nó không chỉ chứa nhiều thông tin hữu ích, việc nghiên cứu tổng hợp dữ liệu dẫn đến việc phát hiện ra các mẫu và chi tiết rất nhỏ mà lẽ ra có thể bị bỏ qua.

Vì vậy, ta có thể thấy tại sao nó có một vai trò to lớn như vậy trong nghiên cứu. Nghiên cứu là tất cả về việc nghiên cứu các mô hình và xu hướng, tiếp theo là đưa ra giả thuyết và chứng minh chúng. Tất cả điều này được hỗ trợ bởi dữ

liệu thích hợp.

Nhìn từ góc độ rộng hơn, phân tích dữ liệu tổng hợp thành hai loại chính. Cụ thể là phân tích dữ liệu định tính và phân tích dữ liệu định lượng. Trong khi cái thứ hai xử lý dữ liệu số, bao gồm các con số, cái thứ hai có ở dạng không phải văn bản. Nó có thể là bất cứ thứ gì như tóm tắt, hình ảnh, biểu tượng,...

Phân tích định lượng liên quan đến bất kỳ loại phân tích nào được thực hiện trên các con số. Từ những kỹ thuật phân tích cơ bản nhất đến những kỹ thuật nâng cao nhất, kỹ thuật phân tích định lượng bao gồm một loạt các kỹ thuật. Cho dù ta cần thực hiện cấp độ nghiên cứu nào, nếu nó dựa trên dữ liệu số, sẽ luôn có các phương pháp phân tích hiệu quả để sử dụng.

Mục tiêu báo cáo

- Giúp cho sinh viên có cái nhìn tổng quát về kiến thức của Số liệu thống kê
- Biết cách xử lý dữ liệu định tính.
- Nhìn nhận chi tiết, đưa ra nhận xét về vấn đề quan tâm được nhắc đến trong báo cáo.

Giới thiệu về bộ dữ liệu

Trong bài báo cáo này, em sẽ sử dụng bộ dữ liệu từ khảo sát của Kaggle về Machine Learning và Data Science để phân tích giới tính và tiềm năng thu nhập trong lĩnh vực công nghệ.

Vấn đề chênh lệch tỉ lệ nam nữ là vấn đề luôn hiện hữu trong suốt quá trình phát triển của ngành công nghệ. Bản thân em cũng luôn băn khoăn rằng với tư cách là nữ, liệu mình có thể duy trì con đường trở thành nhà phân tích dữ liệu trước các định kiến, áp lực trong một môi trường chiếm đa số là nam không. Em sử dụng bộ dữ liệu trên, với giả sử là có tồn tại bất bình đẳng giới trong ngành phân tích dữ liệu, để tìm hiểu xem điều này ảnh hưởng thế nào đến tiềm năng thu nhập của nam và nữ.

Dữ liệu gồm 39 câu hỏi chính và 8 câu hỏi phụ. Câu trả lời cho các câu hỏi trắc nghiệm (chỉ có thể chọn một lựa chọn duy nhất) được ghi lại trong các cột riêng lẻ. Các câu trả lời cho nhiều câu hỏi lựa chọn (có thể chọn nhiều lựa chọn) được chia thành nhiều cột (với một cột cho mỗi lựa chọn trả lời). Bộ khảo sát đã nhận được 20 036 phản hồi.

Hầu hết dữ liệu trong khảo sát này là dữ liệu định tính, vì vậy em chỉ lấy ra 10 câu hỏi để phân tích, các câu hỏi vẫn được đánh số như cũ để dễ dàng xem lại. Nội dung câu hỏi:

- Câu 1: Cho biết tuổi của bạn
- Câu 2: Giới tính của bạn
  - Nam
  - Nữ
  - Phi tính
  - Không muốn nói
  - Muốn tự mô tả
- Câu 3: Bạn đang sinh sống ở quốc gia nào
- Câu 4: Trình độ giáo dục chính quy cao nhất mà bạn đã đạt được hoặc dự định đạt được trong vòng 2 năm tới?
  - Không được học chính thức qua trường trung học
  - Một số học cao đẳng / đại học mà không cần lấy bằng cử nhân
  - Bằng cử nhân
  - Bằng thạc sĩ
  - Bằng tiến sĩ
  - Bằng giáo sư
  - Tôi không muốn trả lời

- Câu 5: Chọn chức danh gần giống nhất với vai trò hiện tại của bạn (hoặc chức danh gần đây nhất nếu đã nghỉ hưu):
  - Nhà phân tích kinh doanh
  - Nhà phân tích dữ liệu
  - Kỹ sư dữ liệu
  - Nhà khoa học dữ liệu
  - DBA / Kỹ sư cơ sở dữ liệu
  - Kỹ sư học máy
  - Giám đốc sản phẩm / dự án
  - Nhà khoa học nghiên cứu
  - Kỹ sư phần mềm
  - Nhà thống kê
  - Sinh viên
  - Hiện không có việc làm
  - Khác
- Câu 6: Bạn đã viết code và/ hoặc lập trình được bao nhiêu năm rồi?
- Bạn muốn giới thiệu một người theo đuổi khoa học dữ liệu học ngôn ngữ lập trình nào trước tiên?
  - Python
  - R
  - SQL
  - C
  - C ++
  - Java

  - Javascript
  - Julia
  - Swift
  - Bash
  - MATLAB
  - Không có
  - Khác

- Câu 20: Quy mô của công ty nơi bạn đang làm việc là bao nhiêu?
- Câu 21: Khoảng bao nhiêu cá nhân chịu trách nhiệm về khối lượng công việc khoa học dữ liệu tại nơi bạn làm việc?
- Câu 24: Mức lương thưởng hàng năm hiện tại của bạn (ước tính $ USD) là bao nhiêu?

Báo cáo này em sẽ chỉ trình bày lý thuyết và bài toán em đặt ra, phần thực hành với R được trình bày ở file riêng.

Chương 1

Quan sát và xử lý dữ liệu

Trình bày trong file thực hành với R.

Chương 2

Ước lượng tham số

2.1 Phân phối mẫu và ước lượng điểm của tham số

2.1.1 Phân phối mẫu của trung bình mẫu $\overline{X}$

Định nghĩa 2.1 Phân phối xác suất của một thống kê được gọi là phân phối mẫu.

2.1.2 Ước lượng điểm của tham số

Khái niệm ước lượng điểm

Một ước lượng điểm của tham số θ của tổng thể là một giá trị số $\widehat{\Theta}$ của thống kê $\widehat{\Theta}$ .Thống kê $\widehat{\Theta}$ được gọi là công cụ ước lượng điểm.

Một số ước lượng điểm thông dụng

Định nghĩa 2.2 Ước lượng $\widehat{\Theta}$ của θ được gọi là ước lượng không chệch của θ nếu

$$E(\widehat{\Theta} = \theta)$$

Nếu ước lượng là có chệch thì

$$E(\widehat{\Theta} \neq \theta)$$

Ước lượng điểm cho kỳ vọng μ

Giả sử X là biến ngẫu nhiên có $E(X) = \mu$ chưa biết, μ được xem là trung bình của tổng thể. Từ X ta lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ kích thước n. Chọn $X = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ làm ước lượng điểm cho kỳ vọng $E(X) = \mu$. Vì $E(X) = \mu$, nên

X là ước lượng không chệch của μ. Khi có một mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$ thì $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ là một ước lượng điểm không chệch của μ. Ước lượng điểm cho phương sai Tương tự với giả sử như trên, X là biến ngẫu nhiên với phương sai tổng thể $V(X) = \sigma^2$ chưa biết. Hiệu chỉnh đại lượng $S^2 = \dfrac{1}{n}\widehat{\Sigma_{i=1}^{n}(X_i - \overline{X})^2}$, ta được S^2 (phương sai hiệu chỉnh mẫu ngẫu nhiên) là ước lượng không chệch cho σ^2.

Khi có mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$ ta tính được các giá trị cụ thể của $\widehat{S^2}$ và S là $\widehat{s^2}$ và s^2:

$$\widehat{s^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2; \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}\widehat{s^2}$$

là các ước lượng điểm của σ^2. Ước lượng điểm cho tỉ lệ Cho p là một tỷ lệ hay xác suất của một sự kiện A trong tổng thể chưa biết. Ta thực hiện n quan sát độc lập và gọi x là số lần xuất hiện A. Khi đó tần suất mẫu

$$\widehat{P} = \frac{x}{n}$$

là ước lượng điểm cho p. Vì $E(\widehat{P}) = p$, nên $\widehat{P}$ là ước lượng không chệch của p.

Sai số tiểu chuẩn của ước lượng

Độ lệch tiêu chuẩn của ước lượng điểm $\widehat{\Theta}$, ký hiệu là $\sigma_{\widehat{\Theta}}$ được gọi là sai số tiêu chuẩn (standard error). Ước lượng điểm của sai số tiêu chuẩn được ký hiệu là $\widehat{\sigma}_{\widehat{\Theta}}$.

Giả sử $W_x = (x_1, x_2, \ldots, x_n)$ là mẫu ngẫu nhiên kích thước n được xây dựng từ biến ngẫu nhiên $X \sim \mathcal{N}(\mu; \sigma^2)$..Khi đó phân phối mẫu của X là $\mathcal{N}(\mu; \dfrac{\sigma^2}{n})$, sai số tiêu chuẩn của X là

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Nếu σ chưa biết, ta thay độ lệch chuẩn mẫu hiệu chỉnh S vào phương trình trên, thì sai số tiêu chuẩn ước lượng của $\overline{X}$ là

$$\sigma_{\overline{X}} = \frac{S}{\sqrt{n}}$$

Bài toán 2.1 Cho tần suất độ tuổi của nữ có vị trí công việc là Data Scientist (Nhà phân tích dữ liệu).

1. Ước lượng điểm cho độ tuổi trung bình của nữ là $\overline{x} =$.

2. Ước lượng điểm không chệch cho độ lệch tiêu chuẩn σ của độ tuổi là độ lệch tiêu chuẩn hiệu chỉnh mẫu $s =$.

3. Sai số tiêu chuẩn của trung bình mẫu là $\sigma_{\overline{x}} = \sigma/\sqrt{n}$, trong đó $n =$ là cỡ mẫu.

4.Ước lượng điểm cho sai số tiêu chuẩn của trung bình mẫu là $\widehat{\sigma_{\overline{x}}} = \dfrac{s}{\sqrt{n}}$

2.1.3 Ước lượng khoảng

Khoảng tin cậy cho kỳ vọng

Ở phần này ta sẽ dùng lý thuyết trong trường hợp mẫu cỡ lớn, phương sai σ^2 chưa biết. Giả sử mẫu ngẫu nhiên $W_x = (x_1, x_2, \ldots, x_n)$ được thành lập từ biến ngẫu nhiên X của tổng thể với kỳ vọng μ và phương sai σ^2 chưa biết. Nếu kích thước mẫu n lớn, thì X có phân phối xấp xỉ phân phối chuẩn với kỳ vọng μµ và phương sai σ^2/n. Do đó, $U = \dfrac{\overline{X} - \mu}{\sigma}\sqrt{n}$ xấp xỉ phân phối chuẩn tắc. Do đó ta có thể sử dụng kết quả của trương hợp phân phối chuẩn, phương sai đã biết để xây dựng khoảng tin cậy cho μ. Tuy nhiên, vì σ chưa biết, khi n đủ lớn (trong thực hành cho phép vận dụng với $n \geq 30$) ta có thể thay σ bởi độ lệch tiêu chuẩn hiệu chỉnh S mà ít làm ảnh hưởng đến phân phối của U. Từ đó ta nhận được kết quả hữu ích sau đây.

Định lý 2.1 (Khoảng tin cậy hai phía) . Khi n đủ lớn thì thống kê

$$\frac{\overline{X} - \mu}{S}\sqrt{n}$$

có phân phối xấp xỉ phân phối chuẩn tắc. Khi đó,

$$\overline{x} - u_{1-\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x} - u_{1+\alpha/2}\frac{s}{\sqrt{n}}$$

là khoảng tin cậy mẫu lớn cho μ với độ tin cậy $1 - \alpha\%$. Khoảng tin cậy trái

$$-\infty < \mu < \mu < \overline{x} + u_{1+\alpha/2}\frac{s}{\sqrt{n}}$$

Khoảng tin cậy phải

$$\overline{x} - u_{1-\alpha/2}\frac{s}{\sqrt{n}} < \mu < +\infty$$

Khoảng tin cậy cho phương sai

Định lý 2.2 Nếu mẫu ngẫu nhiên $W_x = (x_1, x_2, \ldots, x_n)$ kích thước n được xây dựng từ biến ngẫu nhiên X có phân phối chuẩn $\mathcal{N}(\mu; \sigma^2)$ và S^2 là phương sai hiệu chỉnh mẫu ngẫu nhiên thì biến ngẫu nhiên

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$

tuân theo quy luật phân phối khi bình phương với $(n1)$ bậc tự do, ký hiệu là $\chi^2 \sim \chi^2(n-1)$.

Nếu thay $\overline{X}$ bằng μ thì

$$\chi^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n)$$

Vì ta chưa biết thu nhập trung bình của Data Scientist là nữ chưa biết nên ta sẽ sử dụng trường hợp kỳ vọng chưa biết.

Chọn cặp số không âm α_1, α_2 thoả mãn $\alpha_1 + \alpha_2 = \alpha$, tìm các phân vị $\chi^2_{(n-1,\alpha_1)}$ và $\chi^2_{(n-1,1-\alpha_2)}$ thoả mãn $P(\chi^2 < \chi^2_(n-1,\alpha_1)) = \alpha_1; \quad P(\chi^2 < \chi^2_(n-1,1-\alpha_2)) = \alpha_2$.

Từ đó suy ra Khoảng tin cậy hai phía $(\alpha_1 = \alpha_2 = \alpha/2)$

$$\frac{(n-1)s^2}{\chi^2_{(n-1,1-\alpha/2)}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{(n-1,\alpha/2)}}$$

Khoảng tin cậy trái $(\alpha_1 = \alpha, \alpha_2 = 0)$

$$0 < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{(n-1,\alpha/2)}}$$

Khoảng tin cậy phải $(\alpha_1 = 0, \alpha_2 = \alpha)$

$$\frac{(n-1)s^2}{\chi^2_{(n-1,1-\alpha/2)}} < \sigma^2 < +\infty$$

Chú ý 2.1 .

1. Các giá trị $\chi^2_{(n-1,1-\alpha/2)}, \chi^2_{(n-1,\alpha/2)}, \chi^2_{(n-1,\alpha/2}, \chi^2_{(n-1,1-\alpha/2)}$ được tra từ bảng phân phối khi bình phương với $n-1$ bậc tự do.

2. Lấy căn bậc hai các cận trong các khoảng tin cậy cho phương sai ta sẽ thu được các khoảng tin cậy cho các độ lệch chuẩn.

Khoảng tin cậy cho xác suất với mẫu cỡ lớn

Định lý 2.3 Nếu n đủ lớn thì phân phối của

$$U = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\widehat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

xấp xỉ phân phối chuẩn tắc.

Khi có mẫu cụ thể ta tính được giá trị cụ thể $\widehat{p}$ của $\widehat{P}$ và nhận được khoảng tin cậy của tỷ lệ p như sau.

Khoảng tin cậy hai phía (đối xứng)

$$\widehat{p} - u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} < p < \widehat{p} + u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \tag{2.1}$$

Khoảng tin cậy trái

$$0 < p < \widehat{p} + u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \tag{2.2}$$

Khoảng tin cậy phải

$$\widehat{p} - u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} < p < 1 \tag{2.3}$$

Chú ý 2.2 Vì p chưa biết nên ta không kiểm tra được điều kiện $np \geq 5$ và $n(1-p) \geq 5$.Trong thực hành ta dùng các điều kiện $n\widehat{p} \geq 5$ và $n(1-\widehat{p}) \geq 5$

Bài toán 2.2 Sau khảo sát, ta thu được thu nhập của nữ ở vị trí Data Scientist. Bây giờ, ta cần ước lượng khoảng tin cậy của thu nhập trung bình và phương sai σ^2 với độ tin cậy 95%

Bài toán 2.3 Xét tỉ lệ nữ Data Scientist

Bài toán 2.4 Xét ngôn ngữ lập trình Python chiếm bao nhiêu phần trăm trong đề xuất (của cả nam và nữ làm Data Scienctist - Sử dụng Q8)

Xác định kích thước mẫu

Trường hợp ước lượng cho giá trị trung bình

Bài toán 2.5 Giả sử

Ở đây ta chỉ xét X có phân phối chuẩn và phương sai chưa biết, vì vậy

$$P(|\overline{X-\mu}| \neq t_{1-\alpha/2}^{(n-1)}\frac{S}{\sqrt{n}}) = \gamma$$

Do đó nếu n thoả mãn

$$t_{1-\alpha/2}^{(n-1)}\frac{s}{\sqrt{n}} \leq \epsilon$$

hay tương đương với

$$n \geq \frac{s^2(t_{1-\alpha/2}^{(n-1)})^2}{\epsilon^2} \tag{2.4}$$

thì ta sẽ có

$$P(|\overline{X}-\mu| \leq \epsilon) \leq \gamma \tag{2.5}$$

Ta không thể tìm n thoả mãn vì cả s và $t_{1-\alpha/2}^{(n-1)}$ phụ thuộc vào n. Nhìn vào bậc phân phối Student, khi số bậc tự do lớn thì các phân vị của phân phối Student $t_{1-\alpha/2}^{(n-1)}$ và của phân phối chuẩn tắc $u_{1-\alpha/2}$ gần như nhau, vì vậy ta có thể chọn n thoả mãn

$$n \geq \frac{s^2(u_{1-\alpha/2})^2}{\epsilon^2} \tag{2.6}$$

Trường hợp ước lượng cho tỷ lệ

Ta xét một tổng thể mà mỗi cá thể hoặc có tính chất A hoặc không có tính chất A nào đó. Gọi p là tỷ lệ cá thể có tính chất A trong tổng thể. Thông thường p chưa biết. Giả sử trên một mẫu ngẫu nhiên cỡ n có X cá thể có tính chất A. Khi đó tần suất $\widehat{P} = \dfrac{X}{n}$ có phân phối xấp xỉ phân phối chuẩn với trung bình là p và phương sai $\dfrac{p(1-p)}{n}$. Do đó

$$P(|\widehat{P}-p| \leq u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{\sqrt{n}}}) = \gamma \tag{2.7}$$

ở đây $\alpha = 1-\gamma$. Do đó nếu n thoả mãn

$$u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{\sqrt{n}}} \leq \epsilon$$

hay tương đương với

$$n \geq \frac{(u_{1-\alpha/2})^2 p(1-p)}{\epsilon^2} \tag{2.8}$$

thì

$$P\left(|\widehat{P} - p| \leq \epsilon\right)$$

Tuy nhiên, vì giá trị của p chưa biết, nên vế phải của (2.53) chưa xác định. Có hai cách để khắc phục tình trạng này 1. Cách thứ nhất là lấy một mẫu sơ bộ kích thước k để thu được tần suất $\widehat{p_k}$ và lấy $\widehat{p_k}$ làm ước lượng ban đầu cho p. Khi đó bất đẳng thức (2.53) trở thành

$$n \geq \frac{(u_{1-\alpha/2})^2 \widehat{p_k}(1-\widehat{p_k})}{\epsilon^2} \tag{2.9}$$

với điều kiện

$$k\widehat{p_k} \geq 5 \quad \text{và} \quad k(1-\widehat{p_k}) \geq 5 \tag{2.10}$$

2. Cách thứ hai, sử dụng bất đẳng thức Cauchy $p(1-p) \leq \frac{1}{4}$ ta nhận được

$$u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq \frac{u_{1-\alpha/2}}{2\sqrt{n}}$$

Nếu ta chọn n thoả mãn điều kiện

$$n \geq \frac{(u_{1-\alpha/2})^2}{4\epsilon^2} \tag{2.11}$$

Vậy ta sẽ lấy n nguyên dương nhỏ nhất thoả mãn phương trình trên.

Bài toán 2.6 Phải lấy cỡ mẫu bao nhiêu để lương trung bình của nữ Data Scientist là \$139,542 với độ tin cậy.

Bài toán 2.7 Phải lấy cỡ mẫu bao nhiêu để tỉ lệ nữ Data Scientist chiếm 19,7% với độ tin cậy.

Chương 3

Kiểm định giả thuyết thống kê

3.1 Kiểm định giả thuyết về tham số của tổng thể

3.1.1 Kiểm định giả thuyết về kỳ vọng của phân phối chuẩn với phương sai chưa biết

Phương pháp kiểm định với mức ý nghĩa cố định

- Chọn thống kê thử nghiệm (tiêu chuẩn kiểm định)

$$U = \frac{\overline{X} - \mu_0}{S}\sqrt{n} \tag{3.1}$$

  Nếu giả thuyết $H_0 = \mu = \mu_0$ là đúng thì U có phân phối chuẩn tắc $\mathcal{N}(0;1)$

- Xây dựng tiêu chuẩn bác bỏ giả thuyết H_0. Miền bác bỏ giả thuyết H_0, ký hiệu là W_α, được xây dựng phụ thuộc vào thuyết đối H_1.

H_0	H_1	Miền bác bỏ giả thuyết H_0
$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty; -u_{1-\alpha/2}) \bigcup (u_{1-\alpha/2}; +\infty)$
$\mu = \mu_0$	$\mu > \mu_0$	$(u_{1-\alpha}; +\infty)$
$\mu = \mu_0$	$\mu < \mu_0$	$(-\infty; -u_{1-\alpha})$

  trong đó $u^{(n-1)}_{1-\alpha/2}$ và $u^{(n-1)}_{1-\alpha}$ được tra từ bảng phân phối Student.

- Tính giá trị quan sát của tiêu chuẩn kiểm định: Lập mẫu cụ thể $W_X = (X_1, X_2, \ldots, X_n)$ và tính các giá trị $\overline{x}$ và s của mẫu, từ đó giá trị quan sát

được tìm là

$$u_{qs} = \frac{\overline{x} - \mu_0}{s}\sqrt{n} \tag{3.2}$$

- Xét xem t_{qs} có thuộc W_α hay không để kết luận
  - Nếu $u_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0.
  - Nếu $u_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0.

Bài toán 3.1 Kiểm định giả thuyết lương trung bình của nữ Data Scientist là $\mu_0 = 139,542$ với độ tin cậy 95%

Sử dụng ngôn ngữ R

3.1.2 Kiểm định giả thuyết cho phương sai và độ lệch chuẩn của biến ngẫu nhiên phân phối chuẩn

Trường hợp chưa biết kỳ vọng

- Chọn tiêu chuẩn kiểm định

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \tag{3.3}$$

- Xây dựng miền bác bỏ giả thuyết H_0 phụ thuộc vào thuyết đối H_1

H_0	H_1	Miền bác bỏ W_α
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$(-\infty; -\chi^2_{n-1;\alpha/2}) \cup (\chi^2_{n-1;1-\alpha/2}; +\infty)$
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$(\chi^2_{n-1;1-\alpha}; +\infty)$
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$(-\infty; -\chi^2_{n-1;\alpha})$

trong đó $\chi^2_{n-1;\alpha/2}$, $\chi^2_{n-1;1-\alpha/2}$, $\chi^2_{n-1;1-\alpha}$, $\chi^2_{n-1;\alpha}$ được tra từ bảng phân phối khi bình phương.

- Lập mẫu cụ thể $W_X = (X_1, X_2, \ldots, X_n)$, giá trị quan sát được tìm là

$$\chi^2_{qs} = \frac{(n-1)s^2}{\sigma_0^2} \tag{3.4}$$

- Xét xem χ^2_{qs} có thuộc W_α hay không để kết luận.

– Nếu $\chi^2_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0.

– Nếu $\chi^2_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0.

Bài toán 3.2 So sánh kỳ vọng của lương Data Scientist nam và nữ với độ tin cậy 95%

3.1.3 Kiểm định giả thuyết về tỷ lệ của tổng thể

Mẫu cỡ lớn

Điều kiện là $np_0 \geq 5$ và $n(1-p_0) \geq 5$.

- Với giả thuyết H_0 đúng xét thống kê

$$U = \frac{\widehat{P} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \tag{3.5}$$

  Thống kê U tuân theo luật phân phối chuẩn tắc $\mathcal{N}(0;1)$.

- Xây dựng miền bác bỏ giả thuyết H_0 phụ thuộc vào thuyết đối H_1 như sau:

H_0	H_1	Miền bác bỏ W_α
$p = p_0$	$p \neq p_0$	$(-\infty; -u_{1-\alpha/2}) \bigcup (u_{1-\alpha/2}; +\infty)$
$p^2 = p_0$	$p^2 > p_0$	$(u_{1-\alpha}; +\infty)$
$p = p_0$	$p < p_0$	$(-\infty; -u_{1-\alpha})$

- Lập mẫu cụ thể, tính giá trị quan sát của tiêu chuẩn kiểm định

$$u_{qs} = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)}\sqrt{n}} \tag{3.6}$$

- Xét xem u_{qs} có thuộc W_α hay không để kết luận.

  – Nếu $u_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0

  – Nếu $u_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0

Bài toán 3.3

3.1.4 Kiểm định giả thuyết về tỷ lệ của tổng thể

Sai lầm loại II và xác định cỡ mẫu. Giả sử p là tỷ lệ thực tế của cá thể có tính chất A trong tổng thể. Xác suất mắc sai lầm loại II của kiểm định hai phía với thuyết đối $H_1 : p \neq p0$ là

$$\beta = \phi\left(\frac{p_0 - p + u_{1-\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right) - \phi\left(\frac{p_0 - p - u_{1-\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right) \quad (3.7)$$

Nếu $H_1 : p < p_0$ thì

$$\beta = 1 - \phi\left(\frac{p_0 - p - u_{1-\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right) \quad (3.8)$$

Nếu $H_1 : p > p_0$ thì

$$\beta = \phi\left(\frac{p_0 - p + u_{1-\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right) \quad (3.9)$$

Từ các phương trình này ta có thể tìm được kích thước mẫu gần đúng, n, cho phép kiểm tra mức α có xác suất rủi ro β cho trước. Nếu $H_1 : p \neq p_0$ thì

$$n = \left[\frac{u_{1-\alpha/2}\sqrt{p_0(1-p_0)} + u_{1-\beta}\sqrt{p(1-p)}}{p - p_0}\right]^2 \quad (3.10)$$

Nếu $H_1 : p > p_0$ hoặc $H_1 : p < p_0$ thì

$$n = \left[\frac{u_{1-\alpha}\sqrt{p_0(1-p_0)} + u_{1-\beta}\sqrt{p(1-p)}}{p - p_0}\right]^2 \quad (3.11)$$

3.2 So sánh tham số của hai tổng thể

3.2.1 So sánh hai kỳ vọng

Các phương sai chưa biết, $n_1 \geq 30, n_2 \geq 30$

- Chọn tiêu chuẩn kiểm định

$$U = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Nếu giả thuyết H_0 đúng thì $\mu_1 - \mu_2 = 0$,

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{3.12}$$

Vì X và Y độc lập nên U có phân phối chuẩn tắc $\mathcal{N}(0;1)$.

- Miền bác bỏ giả thuyết H_0 được xác định cho ba trường hợp

H_0	H_1	Miền bác bỏ W_α
$\mu_1 - \mu_2 = \Delta_0$	$\mu_1 - \mu_2 \neq \Delta_0$	$(-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}; +\infty)$
$\mu_1 - \mu_2 = \Delta_0$	$\mu_1 - \mu_2 > \Delta_0$	$(u_{1-\alpha}; +\infty)$
$\mu_1 - \mu_2 = \Delta_0$	$\mu_1 - \mu_2 < \Delta_0$	$(-\infty; -u_{1-\alpha})$

- Từ hai mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_{n_1})$, $W_y = (y_1, y_2, \ldots, y_{n_2}$, ta tính được giá trị quan sát của tiêu chuẩn kiểm định:

$$u_{qs} = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.13}$$

Nếu ký hiệu s là độ lệch tiêu chuẩn hiệu chỉnh hợp nhất thì $s = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ khi đó $u_{qs} = \frac{\overline{x} - \overline{y}}{s}$

- Xét xem u_qs có thuộc W_α hay không để kết luận
  - Nếu $u_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0
  - Nếu $u_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0

3.2.2 So sánh hai tỷ lệ

Phương pháp kiểm định với mức ý nghĩa cố định

- Xét $U = \frac{(\widehat{P_1} - \widehat{P_2}) - (p_1 - p_2)}{\sqrt{\widehat{P}(1-\widehat{P})(\frac{1}{n_1} + \frac{1}{n_2})}}$. Nếu giả thuyết H_0 đúng thì $p_1 = p_2$ và tiêu chuẩn kiểm định là

$$U = \frac{(\widehat{P_1} - \widehat{P_2})}{\sqrt{\widehat{P}(1-\widehat{P})(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{3.14}$$

Nếu $(n_1 + n_2)\widehat{P} > 5$ và $(n_1 + n_2)(1 - \widehat{P}) > 5$ thì $U \sim \prime; \infty$

- Miền bác bỏ giả thuyết H_0 được xác định phụ thuộc vào thuyết đối H_1 như sau:

H_0	H_1	Miền bác bỏ W_α
$p_1 = p_2$	$p_1 =_2$	$(-\infty; -u_{1-\alpha/2}) \bigcup (u_{1-\alpha/2}; +\infty)$
$p_1 = p_2$	$p_1 > p_2$	$(u_{1-\alpha}; +\infty)$
$p_1 = p_2$	$p_1 < p_2$	$(-\infty; -u_{1-\alpha})$

- Từ mẫu thu thập, ta tính được giá trị quan sát của tiêu chuẩn kiểm định:

$$u_q s = \frac{(\widehat{p_1} - \widehat{p_2})}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{3.15}$$

với $\widehat{p_1} = \frac{m_1}{n_1}$, $\widehat{p_2} = \frac{m_2}{n_2}$

3.2.3 So sánh hai phương sai

3.3 Phân tích phương sai một nhân tố

- Chọn tiêu chuẩn kiểm định:

$$F = \frac{(n-k)\sum_{i=1}^{k} n_i(\overline{X_i} - \overline{X})^2}{(k-1)\sum_{i=1}^{k}[\sum_{i=1}^{k} n_i(\overline{X_i} - \overline{X})^2]} \tag{3.16}$$

- Với các giả thiết các biến Xi có phân phối chuẩn và có cùng phương sai, với giả thuyết H_0 đúng thì F có phân phối Fisher với bậc tự do là $(k1; nk)$.

- Miền bác bỏ giả thuyết H_0 là $W_\alpha = (F_{1\alpha}(k1; nk); +\infty)$ trong đó $(F_{1\alpha}(k1; nk)$ là phân vị mức 1α của phân phối Fisher với bậc tự do là $(k1; nk)$.

Chương 4

Phân tích tương quan và hồi quy

4.1 Kiểm định về hệ số tương quan tuyến tính

- Đưa ra cặp giả thuyết cần kiểm định
- Chọn tiêu chuẩn kiểm định với H_0 đúng. Nếu giả thiết trong Định lý 1 thỏa mãn thì T có phân phối Student với $n2$ bậc tự do.
- Tìm miền bác bỏ giả thuyết H_0
- Tính giá trị quan sát của tiêu chuẩn kiểm định và kết luận.

4.2 Kiểm định và ước lượng hệ số hồi quy

- Đưa ra cặp giả thuyết cần kiểm định
- Chọn tiêu chuẩn kiểm định

$$T = \frac{\widehat{\beta} - \beta}{\sqrt{\text{MSE}/S_{xx}}} \tag{4.1}$$

  với H_0 đúng. Nếu Giả thiết 1 và Giả thiết 2 được thỏa mãn thì T có phân phối Student với $n2$ bậc tự do.
- Xác định miền bác bỏ W_α
- Tính giá trị quan sát của tiêu chuẩn kiểm định và kết luận.

Kết luận

Nữ trong ngành Khoa học dữ liệu còn hiếm nhưng không phải là không có tiềm năng. Cần phân tích sâu hơn nữa để biết lí do vì sao thiếu hụt nữ trong ngành này.

Tài liệu tham khảo

Tiếng Việt

[1] Nguyễn Thị Thu Thủy (2021), *Suy luận thống kê*, Bài giảng, Bộ môn Toán ứng dụng – Viện Toán ứng dụng và Tin học – Trường Đại học Bách khoa Hà Nội (lưu hành nội bộ).

Tiếng Anh

[2] Kaggle Machine Learning and Data Science 2020