ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC - KỸ THUẬT MÁY TÍNH

xác suất thống kê

Ôn cuối kì

GVHD: Nguyễn Thị Kiều Dung
SV thực hiện: ********** _ *******

Tp. Hồ Chí Minh, Tháng 8/2017

Mục lục

Đây là tài liệu mình soạn cho mục đích ôn tập cá nhân.

Nếu bạn có ghé qua đây. Mình sẽ rất vui nếu nó có ích với bạn. Tuy nhiên, vì mục đích soạn tài liệu này dùng cho mình tự ôn tập nên mình sẽ không chịu trách nhiệm về bất cứ sai lầm nào có trong tài liệu này.

Chúc bạn ôn tập tốt và trân trọng những thứ nhỏ nhặt nhất bên cạnh mình!!!

1 Biến ngẫu nhiên(Random variable)

1.1 Định nghĩa và phân loại

Biến ngẫu nhiên(random variable) là một biến số trong kết quả của mỗi phép thử nó sẽ nhận một và chỉ một trong các giá trị có thể có của nó tùy thuộc vào sự tác động của các yếu tố ngẫu nhiên.

Có 2 loại:

- Biến ngẫu nhiên rời rạc(discrete random variable)
- Biến ngẫu nhiên liên tục(continuous random variable)

1.2 Các phân phối xác suất của biến ngẫu nhiên

1.2.1 Bảng phân phối xác suất

1.2.2 Hàm mật độ xác suất

Ý nghĩa: để biểu thị mức độ tập trung xác suất của biến ngẫu nhiên liên tục trong lân cận của một điểm.

Định nghĩa: $\begin{cases} f(x) \geq 0, \forall x \\ \int_{-\infty}^{+\infty} f(x)dx = 1 \end{cases}$

Tính chất:

- $P(a \leqslant X \leqslant b) = \int_a^b f(x)dx$
- $P(X = x_0) = 0, \forall x_0$
- $P(a \leqslant X < b) = P(a \leqslant X \leqslant b) = P(a < X < b) = P(a < X \leqslant b)$

1.3 Hàm phân phối xác suất(hàm phân bố tích lũy - cumulative distribution function)

Định nghĩa:$F(x) = P(X < x), x \in \mathbb{R}$

Ý nghĩa: phản ảnh mức độ tập trung xác suất của BNN X ở về phía bên trái x_0

Tính chất:

- $0 \leqslant F(x) \leqslant 1, \forall x \in \mathbb{R}$
  $F(-\infty) = 0; F(+\infty) = 1$
- Nếu $x_1 < x_2$ thì $F_{x_1} \leqslant F_{x_2} => F_x$ là hàm tăng trên $\mathbb{R}$
- $P(a \leqslant X < b) = F(b) - F(a)$

1.4 Một số tham số đặc trưng của biến ngẫu nhiên

1.4.1 Kỳ vọng toán

Kỳ vọng toán(Expectation/Mean) của BNN.
Kí hiệu : E(x) hay M(x).
Công thức tính:

- Đối với BNN rời rạc:$E(X) = \sum_i X_i P_i$
- Đối với BNN liên tục: $E(X) = \sum_{-\infty}^{+\infty} x.f(x)dx$

1.4.2 Phương sai và độ lệch chuẩn

Phương sai(variance) bằng trung bình của bình phương sai lệch giữa các biến ngẫu nhiên với kỳ vọng toán của nó.
Kí hiệu: D(x) hay V(x)
Công thức tính: $D(x) = E(x^2) - (E(x))^2$
Đô lệch chuẩn(standard deviation) của biến ngẫu nhiên x, kí hiệu σ_x, là căn bậc hai của phương sai: $\sigma_x = \sigma(X) = \sqrt{D(X)}$

2 Vector ngẫu nhiên

2.1 Đặc trưng của BNN hai chiều

- $E(x, y) = (E(x), E(y)$
- Hiệp phương sai(covarian):
  $cov(X, Y) = E(XY) - E(X)E(Y)$

3 Lý thuyết mẫu

Tổng thể thống kê là tập hợp các phần tử thuộc đối tượng nghiên cứu, cần được quan sát, thu thập và phân tích theo một hoặc một số đặc trưng nào đó.

Các phần tử tạo thành tổng thể thống kê được gọi là *đơn vị tổng thể*.

Mẫu là một số đơn vị được chọn ra từ tổng thể theo một phương pháp lấy mẫu nào đó. Các đặc trưng mẫu được sử dụng để suy rộng ra các đặc trưng của tổng thể nói chung.

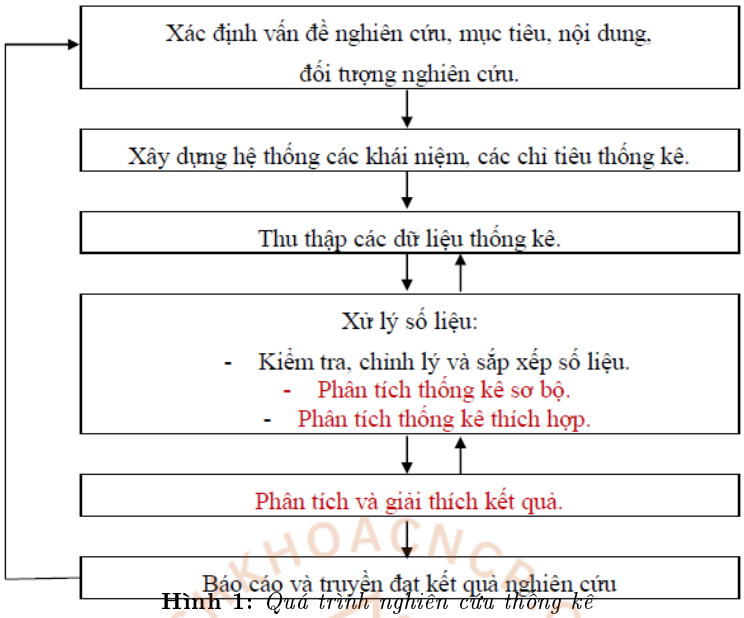
Đặc điểm thống kê(dấu hiệu nghiên cứu) là các tính chất quan trọng liên quan trực tiếp đến nội dung nghiên cứu và khảo sát cần thu nhập dữ liệu trên các đơn vị tổng thể. Có 2 loại đặc điểm thống kê: *đặc điểm thuộc tính* và *đặc điểm số lượng*

Có 2 nhóm kỹ thuật lấy mẫu là *kỹ thuật lấy mẫu xác suất*(probability sampling), trên nguyên tắc mọi phần tử trong tổng thể đều có cơ hội được lấy vào mẫu như nhau và *kỹ thuật lấy mẫu phi xác suất*(non-probability sampling).

3.1 Các kỹ thuật lấy mẫu xác suất(probability sampling)

3.1.1 Lấy mẫu ngẫu nhiên đơn giản(simple random sampling)

cách lấy mẫu:



Hình 1: *Quá trình nghiên cứu thống kê*

- lập danh sách tổng thể theo số thứ tự, gọi là khung lấy mẫu.
- Xác định số lượng phần tử n cần lấy mẫu(sample size)
- Chọn 1 mẫu gồm các đối tượng có số thứ tự được lựa chọn ra 1 cách ngẫu nhiên bằng cách bốc thăm, lấy từ 1 bảng số ngẫu nhiên; bằng MTBT hay 1 phần mềm thống kêk nào đó

Ưu điểm: Tính đại diện cáo
Khuyết điểm: mẫu phải không có kích thước quá lớn; người nghiên cứu phải lập được danh sách tổng thể cần khảo sát.

3.1.2 Lấy mẫu hệ thống(Systematic sampling)

Cách lấy mẫu:

- Lập thành danh sách N phần tử của tổng thể, có mã là số thứ tự
- Xác định số phần tử n cần lấy vào mẫu(sample size)
- Xác định số nguyên K- gọi là khoảng cách, k lấy giá trị làm tròn N/n. Chọn phần tử đầu tiên vào mẫu 1 cách ngẫu nhiên (có số thứ tự trong khoản 1 đến k hay 1 đến N). Các phần tử tiếp theo là các phần tử có STT= STT phần tử đầu tiên +k/2k/3k....

Ưu điểm: tiết kiệm thời gian khi cần mẫu có kích thước lớn.
Khuyết điểm: người nghiên cứu phải lập được danh sách tổng thể cần khảo sát. Thứ tự trong danh sách tổng thể chỉ để mã hóa, không được sắp xếp theo các đặc điểm khảo sát.

3.1.3 Lấy mẫu phân tầng(stratified sampling)

Cách lấy mẫu:

- Chia tổng thể thành nhiều tầng khác nhau dựa và các tính chất liên quan đến đặc điểm cần khảo sát. Trên mỗi tầng thực hiện lấy mẫu ngẫu nhiên đơn giản(simple probability sampling) với số lượng phần tử cần lấy vào mẫu là n_i được phân bổ theo tỉ lệ các phần tử ở mỗi tầng.
- Trong thực tế, với mẫu được chọn, người ta có thể kết hợpk khảo sát thêm các đặc điểm riêng lẻ đối với những phần tử trong cùng 1 tầng. Khi đó nếu nhận thấy 11 vài giá trị m_i quá nhỏ làm các khảo sát riêng lẻ đó không đủ độ tin cậy thì chúng ta cần lấy mẫu không cân đối(disproportionately) và phải quan tâm đến việc hiệu chỉnh kết quả theo trọng số.

Ưu điểm: kỹ thuật này làm tăng khả năng đại diện mẫu theo đặc điểm cần khảo sát. Ở các nghiên cứu có quy mô lớn, người ta thường kết hợp với lấy mẫu cả cụm.

3.1.4 Lấy mẫu cả cụm(cluster sampling) và lấy mẫu nhiều giai đoạn(multi-stage sampling)

Cách lấy mẫu:

- Chia tổng thể thành nhiều cụm theo các tính chất nào đó ít liên quan đến đặc tính cần khảo sát, chọn ra m cụm ngẫu nhiên. Khảo sát hết các phần tử trong các cụm đã lấy ra. Theo các này, số phần tử lấy vào mẫu có thể nhiều hơn số cần thiết n và các phần tử trong cùng cụm có khuynh hướng giống nhau.
- Để kkhắc phục, t chọn m cụm gọi là mẫu bậc 1 nhưng không khảo sát hết mà trong từng cụm bậc 1 lại chọn ngẫu nhiên K_i cụm nhỏ gọi là mẫu bậc 2;.. làm như vậy cho đến khi đủ số lượng cần. Khảo sát tất cả các phần tử đã được chọn ở bậc cuối cùng.

Ưu điểm: kỹ thuật này xử lý tốt các khó khăn gặp phải khi tổng thể có phân bố rộng về mặt địa lý(thời gian, tiền bạc, nhân lực, bảo quản dữ liệu,..) hay hi lập 1 danh sách tổng thể đầu đủ khó khăn.

3.2 Kỹ thuật lấy mẫu phi xác suất(non-probability sampling)

3.2.1 Lấy mẫu thuận tiện(convenient sampling)

Người lấy mẫu lấy thông tin cần khảo sát ở những nơi mà người đó nghĩ là thuận tiên.

3.2.2 Lấy mẫu định mức(quota sampling)

Người lấy mẫu chia tổng thể thành cá tổng thể con(tương tự như phân tầng trong lấy mẫu xác suất) rồi dựa vào kinh nghiệm tự định mức số phần tủ cần lấy theo tỷ lệ nào đó.

3.2.3 Lấy mẫu phán đoán(judgement sampling)

Người lấy mẫu dựa vào năng lực và kinh nghiệm của mình để phán đoán cần khảo sát trong phạm vi nào, những phần tử nào cần chọn vào mẫu.

4 Các đặc trưng tổng thể và mẫu

5 Lý thuyết ước lượng

Có 2 cách ước lượng:

Các đặ trưng của mẫu tổng quát	Các đặc trưng của mẫu cụ thể
Trung bình mẫu: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	Trung bình mẫu: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$
Phương sai mẫu $\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$	phương sai mẫu: $\hat{s}^2$ Độ lệch mẫu: $\hat{s}$ $\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n} n_i x_i^2 - \bar{x}^2 = \bar{x^2} - \bar{x}^2$
Phương sai mẫu hiệu chỉnh $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}\hat{S}^2$	phương sai mẫu hiệu chỉnh: s^2 Độ lệch mẫu hiệu chỉnh s $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\hat{s}^2$
Tỉ lệ mẫu $F = \frac{M}{N}$	Tỉ lệ mẫu $f = \frac{m}{n}$

- *Ước lượng điểm* : là dùng một tham số thống kê mẫu đơn lẻ để ước lượng giá trị tham số của tổng thể. Ví dụ dùng một giá trị cụ thể của trung bình mẫu $\bar{X}$ để ước lượng trun bình tổng thể a.
- *Ước lượng khoảng* : là tìm ra khoảng ước lượng $(G_1; G_2)$ cho tham số θ trong tổng thể sao cho ứng với độ tin cậy(confidence) bằng(1-α) cho trước, P($G_1 < \theta < G_2$)=1-α.

Tham số cần ước lượng	Phân bố của tổng thể	Thông tin bổ sung	Khoảng tin cậy khi chọn $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$
Tỉ lệ P(xác suất)	Nhị thức B(1,p)	Mẫu lớn ($n \geq 30$)	$(F \pm Z_\alpha \frac{\sqrt{f(1-f)}}{\sqrt{n}})$
Trung bình a	Bất kỳ	mẫu lớn($n \geq 30$)	$(\overline{X} \pm Z_\alpha \frac{s}{\sqrt{n}})$
	Chuẩn N(a,σ^2)	σ^2 đã biết	$\overline{X} \pm Z_\alpha \frac{\sigma}{\sqrt{n}}$
	Chuẩn N(a,σ^2)	σ^2 chưa biết, mẫu nhỏ (n<30)	$\overline{X} \pm Z_\alpha \frac{s}{\sqrt{n}}$
Phương sai σ^2	chuẩn N(a,σ^2)	a chưa biết	$(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)})$

Lưu ý:

- Tìm giá trị Z_α: tra ngược bảng tích phân Laplace
- Tìm giá trị $T^{n-1}_{\frac{\alpha}{2}}$: tra bảng student, cột $\frac{\alpha}{2}$, dòng n-1
- Tìm giá trị $\chi^2_{\frac{\alpha}{2}}(n-1)$:tra bảng chi bình phương, cột $\frac{\alpha}{2}$, dòng n-1.

6 Kiểm định giả thuyết

Giả thiết kiểm định H_0:

- Giả thiết về tham số của tổng thể
- Giả thuyết về dạng phân phối của tổng thể
- Giả thuyết về tính độc lập của các biến ngẫu nhiên

Giả thuyết H_1 là một mệnh đề mâu thuẫn với H_0, H_1 thể hiện xu hướng cần kiểm định.

Tiêu chuẩn kiểm định là hàm thống kê $G = G(X_1, X_2, ...X_n, \sigma_0)$,xây dựng trên mẫu ngẫu nhiên $W = (X_1, X_2, ...X_n)$ va tham số σ_0 liên quan đến H_0; Điều kiên đặt ra với thống kê G là nếu H_0 đúng thì quy luật phân phối xác suất của G phải hoàn toàn xác định.

Miền bác bỏ giả thiết W_α là miền thỏa $P(G \in W_\alpha / H_0$đúng)=α. α là một số khá bé, thường không quá 0.05 và gọi là mức ý nghĩa của kiểm định. Có vô số miền W_α như vậy.

Quy tắc kiểm định: Từ mẫu thực nghiệm, ta tính được một giá trị cụ thể của tiêu chuẩn kiểm định là thống kê $g_{qs} = G(X_1, X_2, ..., X_n, \sigma_0)$. Theo nguyên lý xác suất bé, biến cố $G \in W_\alpha$ có xác suất nhỏ nên với 1 mẫu thực nghiệm, nó không thể xảy ra. Do đó:

- Nếu $g_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận giả thiết H_1
- Nếu $g_{qs} \notin W_\alpha$ thì bác bỏ H_0, thừa nhận giả thiết H_1

hoặc:

- $Z_{qs} \in W_\alpha$ thì bác bỏ H_0
- $Z_{qs} \notin W_\alpha$ thì bác bỏ H_1

1. $W_\alpha = (-\infty, -Z_\alpha) \cup (Z_\alpha, +\infty)$
   $\phi(Z_\alpha) = \frac{1-\alpha}{2}$
2. $W_\alpha = (-\infty, -Z_{2\alpha})$
   $\phi(Z_{2\alpha}) = \frac{1-2\alpha}{2}$
3. $W_\alpha = (Z_{2\alpha}, +\infty)$
   $\phi(Z_{2\alpha}) = \frac{1-2\alpha}{2}$

6.1 Bài toán kiểm định tỉ lệ

	Giả thiết Kiểm định H_0	Giả thiết Kiểm định H_1	Tiêu chuẩn kiểm định	Miền bác bỏ H_0 với mức ý nghĩa α
Bài tập 1 mẫu($n \geq 30$)	$p = p_0$	$p \neq p_0$	$z_{qs} = \frac{F-p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$	$W_\alpha = (-\infty, -Z_\alpha) \cup (Z_\alpha, +\infty)$
		$p < p_0$		$W_\alpha = (-\infty, -z_{2\alpha})$
		$p > p_0$		$W_\alpha = (z_{2\alpha}, +\infty)$
Bài tập 2 mẫu $n_1, n_2 \geq 30$	$p_1 = p_2$	$p_1 \neq p_2$	$Z_{qs} = \frac{F_1-F_2}{\sqrt{\overline{f}(1-\overline{f})(\frac{1}{n_1}+\frac{1}{n_2})}}$ $\overline{f} = \frac{n_1F_1+n_2F_2}{n_1+n_2}$	$W_\alpha = (-\infty, -Z_\alpha) \cup (Z_\alpha, +\infty)$
		$p_1 < p_2$		$W_\alpha = (-\infty, -z_{2\alpha})$
		$p_1 > p_2$		$W_\alpha = (z_{2\alpha}, +\infty)$

ở BT 2 mẫu:
$f_1 = \frac{m_1}{n_1}; f_2 = \frac{m_2}{n_2} => \overline{f} = \frac{m_1+m_2}{n_1+n_2}$

6.2 Bài toán kiểm định trung bình

	GTKĐ H_0	GT đối H_1	Tiêu chuẩn kiểm định	miền bác bỏ H_0	
BT 1 mẫu				-phân phối chuẩn, đã biếtσ^2 -n$\geq$30	-phân phối chuẩn, chưa biếtσ^2 -n$<$30
	a=a_0	$a \neq a_0$	$Z_{qs} = \frac{\overline{X}-a_0}{\sigma}\sqrt{n}$ Nếu không có σ^2 thì thay bằng S	$W_\alpha = (-\infty, -Z_\alpha) \cup (Z_\alpha, +\infty)$	$W_\alpha = (-\infty, -t_{\frac{\alpha}{2}}(n-1)) \cup (t_{\frac{\alpha}{2}}(n-1), +\infty)$
		a$<a_0$		$W_\alpha = (-\infty, -Z_{2\alpha})$	$W_\alpha = (-\infty, -t_\alpha(n-1))$
		a$>a_0$		$W_\alpha = (Z_{2\alpha}, +\infty)$	$W_\alpha = (t_\alpha(n-1), +\infty)$

6.3 Bài toán kiểm định phương sai

	Giả thiết KĐ H_0	Giả thiết Đối H_1	ĐK của pp tổng thể	Tiêu chuẩn kiểm định	miền bác bỏ H_0
BT 1 mẫu	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	-Bất kỳ khi khi mẫu lớn -PP chuẩn khi n nhỏ	$\chi^2_{qs} = \frac{(n-1)S^2}{\sigma_0^2}$	$W_\alpha = [0, \chi^2_{1-\frac{\alpha}{2}}(n-1)) \cup (\chi^2_{\frac{\alpha}{2}}(n-1), +\infty)$
		$\sigma^2 < \sigma_0^2$			$W_\alpha = [0, \chi^2_{1-\alpha}(n-1))$
		$\sigma^2 > \sigma_0^2$			$W_\alpha = (\chi^2_{\frac{\alpha}{2}}(n-1), +\infty)$

6.4 Bài toán kiểm định tính độc lập

- Đặt giả thuyết:
  - H_0: x,y độc lập

  - H_1: x,y không độc lập

- $W_\alpha = (\chi^2_\alpha$(số hàng-1)(số cột-1); $+\infty)$
- tỉnh bảng $E_{i,j}$
  $E_{i,j} = \frac{tonghangi * tongcotj}{kich.thuoc.mau}$
- $\chi^2_{qs} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
- Nếu $\chi^2_{qs} \in W_\alpha$thì bác bỏ H_0.
  Ngược lại, bác bỏ H_1

6.5 Kiểm đinh phân phối chuẩn

- Đặt giả thuyết kiểm định:
  - H_0 : mẫu phù hợp với phân phối chuẩn
  - H_1 : mẫu không phù hợp với phân phối chuẩn
- tìm các đặc trưng mẫu $n, \overline{x}, \hat{s}$
  $\overline{x}$ là ước lượng hợp lý cực đại cho a $=> a = \overline{x}$
  $\hat{s}^2$ là ước lượng hợp lý cực đại cho $\sigma^2 => \sigma = \hat{s}$
- $W_\alpha = (\chi^2_\alpha(k - r - 1); +\infty)$

Khoảng (α, β)	$n_i = o_i$	$p_i = p(\alpha < X < \beta) = \Phi(\frac{\beta - a}{\sigma}) - \phi(\frac{\alpha - a}{\sigma})$
$(\infty; 15)$	25	$\Phi(\frac{15 - a}{\sigma}) - (-0.5)$
...	...	...
$(65; +\infty)$	18	$0.5 - \phi(\frac{65 - a}{\sigma})$

- $\chi^2_{qs} = \frac{1}{n} \sum_i \frac{n_i^2}{p_i} - n$

6.6 Phân phối poisson

Tài liệu

[Giáo trình] Nguyễn Đình Huy, Đậu Thế Cấp, Lê Xuân Đại *Giáo trình Xác suất và thống kê* Nhà xuất bản Đại Học Quốc Gia TP.HCM

[Slide] Nguyễn kiều Dung *slide bài giảng xác suất thống kê* Đại Học Bách Khoa TPHCM