3.10pt

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Tạ Công Sơn

Phần II. Thống kê

Xác suất, thống kê

Hà Nội, 2014

Giới thiệu

Thống kê là bộ môn khoa học nghiên cứu quy luật của các hiện tượng ngẫu nhiên có tính chất số lớn trên cơ sở thu thập và xử lý số liệu thống kê các kết quả quan sát về sự vật hiện tượng ngẫu nhiên này.
Nếu ta thu thập được tất cả các số liệu liên quan đến đối tượng cần nghiên cứu thì ta có thể biết về đối tượng đó. Tuy nhiên trong thực thế điều đó không thể thực hiện được vì quy mô của đối tượng nghiên cứu quá lớn hoặc trong quá trình nghiên cứu đối tượng nghiên cứu bị phá hủy. Vì vậy chúng ta cần lấy mẫu để nghiên cứu. Từ mẫu đó ta cần phải đưa ra những kết luận về đối tượng cần nghiên cứu. Đó là nhiệm vụ chính của phần thông kê này.

Nội Dung

Chương 1. Lý thuyết mẫu.
Chương 2. Bài toán ước lượng tham số.
Chương 3. Bài toán kiểm định giả thiết.
Chương 4. Bài toán tương quan và hồi quy.

1. Các khái niệm cơ bản

Một tổng thể là một tập hợp (X) các đối tượng có chung một tính chất nào đó mà ta đang quan tâm. Mỗi một phần tử của tổng thể gọi là cá thể.
Việc chọn ra từ tổng thể một tập con nào đó gọi là **phép lấy mẫu**.
Tập con $(X_1, X_2, ..., X_n)$ này gọi là một **mẫu** .
n gọi là **cỡ mẫu**.
Gọi x_i là giá trị của X_i, khi đó $(x_1, x_2, ..., x_n)$ gọi là **giá trị của mẫu** (hoặc là một thể hiện của mẫu).
Ta nói rằng một mẫu là **mẫu ngẫu nhiên** nếu trong phép lấy mẫu đó mỗi phần tử của tổng thể đều được chọn một cách độc lập và có xác suất được chọn là như nhau.

2. Chọn mẫu và rút gọn mẫu

2.1.Các phương pháp chọn mẫu

(Xem tài liệu)

2.2. Các rút gọn mẫu

1. Rút gọn dạng thu gọn
2. Rút gọn dạng khoảng (ghép lớp).

Ví Dụ 1

Để đánh giá mức thu nhập của các công nhân trong một khu công nghiệp, người ta đã điều tra ngẫu nhiên 50 công nhân. Kết quả thu nhập của từng người (đơn vị là triệu đồng) theo từng tháng như sau:
4; 4.5; 4.2; 3.8; 4; 4.8; 5; 4.5; 4.2; 3.5; 5; 4.8; 4.5; 3.8; 4; 3.8; 3.5; 5; 4.2; 4.5; 3.8; 3.5; 5.2; 5; 4.5; 5.5; 5; 4.8; 4.5; 3.8; 4; 4.2; 4.5; 5; 5.2; 4.8; 4.8; 5; 3.5; 3.8; 3.5; 4.5; 4.2; 5; 4.5; 4.8; 5; 3.8; 3.5; 4.
a. Hãy rút gọn mẫu trên dưới dạng thu gọn.
b. Hãy thu gọn mẫu trên dưới dạng khoảng độ dài mỗi khoảng là 0.5 (triệu).

3.1. Kỳ vọng mẫu

a. Định nghĩa: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

b. Tính chất: $E\overline{X} = \mu$

3.2. Phương sai mẫu

a. Định nghĩa:

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

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

b. Tính chất:

$$E\sigma^2 = \frac{n-1}{n}\sigma_X^2.$$

$$Es^2 = \sigma_X^2.$$

Máy FX500MS và FX570MS

Mở chương trình: Mode→ 2. (Với FX500MS) Mode→Mode→ 1. (Với FX570MS)
Nhập số liệu: $x_1 \rightarrow$ SHIPFT→;→ $m_1 \rightarrow$DT→...$x_k \rightarrow$ SHIPFT→;→ $m_k \rightarrow$DT
Kết quả:
SHIFT→S-Var→1→= cho $\overline{X} =$
SHIFT→S-Var→2→= cho $\sigma =$
SHIFT→S-Var→3→= cho $s =$

Máy FX570ES

Mở chương trình: SHIFT→setup→ $\bigtriangledown \rightarrow 4 \rightarrow 1$
Nhập số liệu: Mode $\rightarrow 3 \rightarrow 1$. Sau đó nhập vào bảng trên màn hình.→ AC
Kết quả:
SHIFT→ $1 \rightarrow 5(4)$(Chọn Var) $\rightarrow 2 \rightarrow =$ cho $\overline{X} =$
SHIFT→ $1 \rightarrow 5(4) \rightarrow 3 \rightarrow =$ cho $\sigma =$
SHIFT→ $1 \rightarrow 5(4) \rightarrow 4 \rightarrow =$ cho $s =$

Ví dụ 2

Với số liệu của ví dụ 1. Hãy tìm trung bình mẫu và các phương sai mẫu.

Ví dụ 3

Số liệu về lượng nước tiêu thụ (m^3/tháng) của 100 hộ gia đình tại huyện X như sau:

Lượng nước tiêu thụ (m^3/tháng)	Tần số
$[0; 25)$	12
$[25; 50)$	33
$[50; 75)$	40
$[75; 100)$	10
$[100; 125)$	5

Hãy tính trung bình mẫu và các phương sai mẫu.

I. Ước lượng điểm

Tổng thể X chưa biết và tham số θ liên quan tới X là tham số ta cần ước lượng. Vấn đề là căn cứ vào n giá trị $x_1, ..., x_n$ của X đo được trên một cỡ mẫu kích thước n lấy ra từ tập hợp chính, ta cần tìm một giá trị gần đúng θ^* của θ.

1. Định nghĩa: **Ước lượng điểm** của θ là một hàm $\theta^* = T(X_1, ..., X_n)$ của mẫu $(X_1, ..., X_n)$.

Ước lượng điểm θ^* của θ được gọi là **ước lượng không chệch** của θ nếu

$$E\theta^* = \theta$$

Ước lượng điểm θ^* của θ được gọi là **ước lượng chệch** của θ nếu

$$E\theta^* = \theta + C$$

C được gọi là độ chệch.

2. Ước lượng điểm của một số tham số quan trọng.

+ Ước lượng điểm cho trung bình là $\overline{X}$, và đó là ước lượng không chệch.

+Ước lượng điểm cho phương sai là s^2 (là ước lượng không chệch), hoặc σ^2 (là ước lượng chệch với độ chệch là $-DX/n$.)

+ Ước lượng điểm cho độ lệch tiêu chuẩn là s (là ước lượng không chệch).

+Ước lượng điểm cho xác suất là $p^* = m/n$ (m là giá trị mẫu thuộc vào tập đang xét.)

+Ước lượng cho trung vị (Median): là median mẫu, k/hiệu là Med và x/đ như sau:

a. Số liệu dạng thu gọn: G/s số liệu xếp theo thứ tự tăng dần thì

Med=$x_{(n+1)/2}$ nếu n lẻ.

Med=$\frac{1}{2}(x_{n/2} + x_{n/2+1})$ nếu n chẵn

b. Số liệu thu gon dạng khoảng: gọi l sao cho $\sum_{i=1}^{l-1} m_i \leq \frac{n}{2} < \sum_{i=1}^{l} m_i$ khi đó

Med=$\frac{n/2 - \sum_{i=1}^{l-1} m_i}{m_l}(x_{l+1} - x_l) + x_l$

Ví dụ 4

Với số liệu của Ví dụ 1. Hãy tìm các ước lượng cho EX, DX, σ_X, Median, và $p = P(X \leq 4.2)$.

Ví dụ 5

Tiến hành đo chiều cao của 100 học sinh lớp 3 ở một số trường trung tiểu học ở một huyện, ta thu đươc kết quả như sau

Khoảng chiều cao (cm)	số em
$[110; 112)$	5
$[112; 114)$	8
$[114; 116)$	14
$[116; 118)$	17
$[118; 120)$	20
$[120; 122)$	16
$[122; 124)$	10
$[124; 126)$	6
$[126; 128)$	4

Ví dụ 5 (tiếp)

a. Hãy ước lượng chiều cao trung bình của trẻ em lớp 3 của huyện.
b. Hãy ước lượng cho bình phương độ tản mát của chiều cao của các em học sinh lớp 3 của huyện.
c. Hãy ước lượng cho độ lệch tiêu chuẩn của chiều cao của các em học sinh lớp 3 của huyện.
d. Hãy ước lượng cho tỉ lệ học sinh có chiều cao từ 116(cm) tới 124(cm)
e. Hãy ước lượng giá trị median của chiều cao của các học sinh lớp 3 của huyện.

1. Định nghĩa

Một khoảng với hai đầu mút $\theta_1^* = \theta_1^*(X_1, ..., X_n)$ và $\theta_2^* = \theta_2^*(X_1, ..., X_n)$ được gọi là **khoảng tin cậy** (khoảng ước lượng) cho tham số θ với **độ tin cậy** (với xác suất) $1 - \alpha$ nếu

$$P(\theta_1^* < \theta < \theta_2^*) = 1 - \alpha$$

Khi đó $\theta_2^* - \theta_1^*$ gọi là độ chính xác của ước lượng.

Công thức

a. Nếu phương sai DX=σ_X^2 đã biết, X có phân bố chuẩn hoặc cỡ mẫu đủ lớn ($n \geq 30$) khi đó khoảng tin cậy của EX là

$$\mu \in \left(\overline{X} - z(\frac{\alpha}{2})\frac{\sigma_X}{\sqrt{n}}; \overline{X} + z(\frac{\alpha}{2})\frac{\sigma_X}{\sqrt{n}}\right)$$

b. Nếu phương sai DX chưa biết, X có phân bố chuẩn khi đó khoảng tin cậy cho EX là

$$\mu \in \left(\overline{X} - t_{n-1}(\frac{\alpha}{2})\frac{s}{\sqrt{n}}; \overline{X} + t_{n-1}(\frac{\alpha}{2})\frac{s}{\sqrt{n}}\right)$$

c. Nếu phương sai DX chưa biết, X chưa biết có phân bố chuẩn nhưng cỡ mẫu đủ lớn ($n \geq 30$) khi đó khoảng tin cậy của EX là

$$\mu \in \left(\overline{X} - z(\frac{\alpha}{2})\frac{s}{\sqrt{n}}; \overline{X} + z(\frac{\alpha}{2})\frac{s}{\sqrt{n}}\right)$$

Ví dụ 6.

Để nghiên cứu tuổi thọ của một dân tộc thiểu số, người ta thống kê tuổi thọ của những người đã mất của dân tộc đó trong năm qua ở các vùng miền khác nhau trên cả nước có dân tộc đó sinh sống. Kết quả như sau

Tuổi thọ (năm)	≤ 3	$(3, 10]$	$(10, 20]$	$(20, 30]$	$(30, 40]$	$(40, 50]$
Số người	15	8	4	3	2	5

$(50, 60]$	$(60, 70]$	> 70
20	18	5

a. Hãy ước lượng tuổi thọ trung bình của dân tộc đó.
b. Với độ tin cậy 95% tuổi thọ trung bình của dân tộc đó thuộc khoảng tin cậy nào nào?
c. Với xác suất 90% có thể nói tuổi thọ trung bình của dân tộc đó cao nhất là bao nhiêu tuổi.

Ví dụ 7.

Để xác định chiều cao trung bình của các cây bạch đàn trong khu rừng rộng trồng bạch đàn ta không đủ điều kiện đo chiều cao của mọi cây trong khu rừng, do đó người ta đo ngẫu nhiên 35 cây. Kết quả như sau

C.cao (m)	6.50 – 7.0	7.0 – 7.5	7.5 – 8.0	8.0 – 8.5	8.5 – 9.0	9.0 – 9.5
Số cây	2	4	10	11	5	3

Với xác suất 95 % ta có thể nói chiều cao trung bình của cây bạch đàn thuộc khu rừng trên nằm trong khoảng nào. Biết rằng chiều cao của các cây bạch đàn tuân theo phân bố chuẩn.

Công thức

$$p \in \left(p^* - z(\frac{\alpha}{2})\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}}, p^* + z(\frac{\alpha}{2})\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}}\right)$$

Ví dụ 8.

Với số liệu của ví dụ 6.
a. Hãy ước lượng tỷ lệ người có tuổi thọ trên 60 tuổi của dân tộc này. Với xác suất 98% tỷ lệ thấp nhất là bao nhiêu?
b. Hãy ước lượng tỷ lệ trẻ em ≤ 3 tuổi bị chết (trong số những người bị chết). Với xác suất 98% tỷ lệ này thuộc khoảng nào?

Ví dụ 9.

Để xác định tỷ lệ nảy mầm của một lô hạt giống ta gieo thử 300 hạt, thấy có 276 hạt nảy mầm. Với độ tin cậy 95% ta có thể nói tỷ lệ nảy mầm của lô hạt tối đa là bao nhiêu?

4. Ước lượng khoảng cho phương sai của biến ngẫu nhiên chuẩn

Công thức

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

Ví dụ 10.

Để khảo sát độ chính xác của một dụng đo độ dài người ta đo trên cùng một mục tiêu 30 lần bằng dụng cụ đó, kết quả nhận được $s^2 = 0.05$. Hãy tìm ước lượng khoảng cho độ chính xác của dụng cụ với độ tin cậy 95%.

Công thức

G/s X, Y là 2 biến ngẫu nhiên với giá trị trung bình μ_1, μ_2 chưa biết, còn phương sai σ_1^2, σ_2^2 đã biết. Gọi $D = \mu_1 - \mu_2$. Với các mẫu ngẫu nhiên của X, Y là $(X_1, ..., X_{n_1}), (Y_1, ..., Y_{n_2})$. Khi đó

$$D \in \left(\overline{X} - \overline{Y} - z(\frac{\alpha}{2})\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; \overline{X} - \overline{Y} + z(\frac{\alpha}{2})\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Ví dụ 11.

Lấy 100 quả trứng từ lô trứng do nhóm gà A đẻ ra, xác định được trọng lượng trứng trung bình là 40 gam. Lấy 120 quả từ lô trứng do nhóm gà B đẻ ra, xác định được trọng lượng trung bình là 44 gam.
Với $\alpha = 5\%$, sự sai khác giữa hai loại trứng gà nằm trong khoảng nào? Biết trọng lượng quả trứng gà là tuân theo phân bố chuẩn với $\sigma_1^2 = \sigma_2^2 = 15$.

6. Độ chính xác của ước lượng và số quan sát cần thiết

Khái niệm

Ước lượng khoảng dạng $(\theta^* - b(n), \theta^* + b(n))$ thì giá trị $b(n)$ gọi là độ chính sác của ước lượng.
Với độ tin cậy cho trước, giá trị ε cho trước, số quan sát n nhỏ nhất sao cho $b(n) \leq \varepsilon$ thì n gọi là số quan sát cần thiết nhận được ước lượng với độ tin cậy và độ chính xác đã cho.

Ví dụ 12.

Quay về ví dụ 7.
a. Độ chính xác của ước lượng nhận được là bao nhiêu?
b. Bây giờ ta muốn độ chính xác của ước lượng là $\varepsilon = 0.1$ thì cần phải đo bao nhiêu cây?

I. Khái niệm cơ bản

Trong chương này chúng ta đề cập đến vấn đề quan trọng của thống kê: Đó là vấn đề kiểm định giả thiết thống kê, nội dung của bài toán như sau.
Căn cứ vào số liệu thu được (mẫu thu đươc) hãy cho một kết luận về một giả thiết thống kê nào đó mà ta đang quan tâm.
Một giả thiết thông kê là một giả thiết về phân bố, hoặc của các tham số đặc trưng của tập hợp đang xét.
Giả thiết đối chứng gọi là đối thiết. Và ta cần phải lựa chọn hoặc đối thiết(K) hoặc giả thiết(H).
Để giải quyết bài toán trên, thông tin duy nhất mà chúng ta có là một mẫu ngẫu nhiên. Vận dụng các kết quả của lý thuyết xác suất ta sẽ tìm một miền S, sao cho khi mẫu $(X_1, ..., X_n) \in S$ thì ta bác bỏ giả thiết H, còn khi $(X_1, ..., X_n) \notin S$ thì ta chấp nhận H.

Công thức

a. Nếu phương sai DX=σ^2 đã biết, X phân bố chuẩn hoặc $n \geq 30$, đặt $z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$

$$S_1 = \{|z| \geq z(\frac{\alpha}{2})\} \;\; S_2 = \{z \geq z(\alpha)\} \;\; S_3 = \{z \leq -z(\alpha)\}$$

b. Nếu phương sai DX chưa biết, X có phân bố chuẩn, đặt $t = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$

$$S_1 = \{|t| \geq t_{n-1}(\frac{\alpha}{2})\} \;\; S_2 = \{t \geq t_{n-1}(\alpha)\} \;\; S_3 = \{t \leq -t_{n-1}(\alpha)\}$$

c. Nếu phương sai DX chưa biết, X chưa biết có phân bố chuẩn nhưng $n \geq 30$ đặt $z = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$

$$S_1 = \{|z| \geq z(\frac{\alpha}{2})\} \;\; S_2 = \{z \geq z(\alpha)\} \;\; S_3 = \{z \leq -z(\alpha)\}$$

Ví dụ 13

Để xác định chiều cao của các em lứa tuổi lên 10 ở nông thôn vùng đồng bằng Bắc bộ người ta lấy ra một mẫu đại diện với các kết quả như sau

C.cao (cm)	< 130	$130 - 135$	$135 - 140$	$140 - 145$	≥ 145
Số em	5	15	30	20	5

G/s chiều cao tuân theo phân bố chuẩn, $DX = 9$. Có thể kết luận chiều cao trung bình của các em lứa tuổi lên 10 ở nông thôn vùng ĐBBB có chiều cao lớn hơn 137 cm; cao hơn 137.5 cm hay không (với mức ý nghĩa $\alpha = 0.05$)?

Ví dụ 14

Một vường phi lao có chiều cao trung bình chưa xác định. Theo hợp đồng đã ký giữa người sản xuất cây con và lâm trường trông cây thì khi nào chiều cao đạt trên 1m mới đem ra trồng để đạt tỷ lệ sống cao.
Người ta điều tra ngẫu nhiên 50 cây trong vườn ươm tính được trung bình $\overline{X} = 1.1$m, $s \approx 0.12$. Hỏi cây phi lao ở vườn ươm đã đem ra trồng được chưa?

Công thức

$$S_1 = \{|z| \geq z(\frac{\alpha}{2})\} \;\; S_2 = \{z \geq z(\alpha)\} \;\; S_3 = \{z \leq -z(\alpha)\}$$

với $z = \dfrac{m/n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$

Ví dụ 15

Một kho hạt giống có tỉ lệ nảy mầm xác định $p_o = 0.9$. Ngẫu nhiên có một thiết bị bị hỏng làm thay đôi điều kiện bên trong kho. Hỏi tỷ lệ nãy mầm của kho hạt giống có bị giảm xuống không (với $\alpha = 0.05$)?.
Để có thông tin về tỷ lệ nãy mầm mới của kho ta làm thí nghiệm 200 hạt thấy có 140 hạt nảy mầm.

Công thức

a. Nếu phương sai σ_X^2, σ_Y^2 đã biết, X, Y phân bố chuẩn hoặc $n_1, n_2 \geq 30$, đặt

$$z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \text{ thì } S_1 = \{|z| \geq z(\frac{\alpha}{2})\} \ S_2 = \{z \geq z(\alpha)\} \ S_3 = \{z \leq -z(\alpha)\}$$

b. Nếu phương sai DX, DY chưa biết, X, Y có phân bố chuẩn và biết hai phương sai bằng nhau, đặt $t = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\dfrac{n_1+n_2}{n_1 n_2}}}$,

$$S_1 = \{|t| \geq t_{n_1+n_2-2}(\frac{\alpha}{2})\} \ S_2 = \{t \geq t_{n_1+n_2-2}(\alpha)\} \ S_3 = \{t \leq -t_{n_1+n_2-2}(\alpha)\}$$

Công thức

c. Nếu phương sai DX, DY chưa biết, X, Y chưa biết có phân bố chuẩn nhưng $n_1, n_2 \geq 30$,

$$z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}; S_1 = \{|z| \geq z(\frac{\alpha}{2})\} \ S_2 = \{z \geq z(\alpha)\} \ S_3 = \{z \leq -z(\alpha)\}$$

Ví dụ 16

Người ta thí nghiệm hai phương pháp chăn nuôi gà khác nhau, sau một tháng, kết quả tăng trọng như sau
Phương pháp I: $n_1 = 100$ con; $\overline{X} = 1.1$ kg, $s_1^2 = 0.04$
Phương pháp II: $n_2 = 150$ con; $\overline{Y} = 1.2$ kg, $s_2^2 = 0.09$ Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận phương pháp II hiệu quả hơn phương pháp I hay không? Giả thiết mức tăng trọng lượng của gà tuân theo phân bố chuẩn và $\sigma_1^2 = \sigma_2^2$.

Ví dụ 17

Nghiên cứu trọng lượng của trẻ em lứa tuổi lên 10 ở thành phố và nông thôn, ta có hai mẫu đại diện sau

kg	< 35	$35 - 38$	$38 - 41$	$41 - 44$	$44 - 47$	$47 - 50$	$50 - 53$	≥ 53
TP	0	2	8	13	20	15	12	8
NT	5	10	12	15	10	3	0	0

a. Với mức ý nghĩa $\alpha = 0,05$ có thể nói trọng lượng trung bình của trẻ em lứa tuổi lên 10 ở hai vùng trên như nhau hay không? hay vùng nào cao hơn?

b. Ở lứa tuổi này ta xem trọng lượng ≥ 50 kg là diện thừa cân, có nguy cơ béo phì. Từ số liệu trên có thể kết luận:

- Tỷ lệ trẻ em lên 10 ở thành phố có nguy cơ béo phì cao hơn 25% hay không (với mức ý nghĩa $\alpha = 0.1$)?

-Tỷ lệ này thấp nhất là bao nhiêu % (với độ tin cậy 0.9)?

Tiêu chuẩn Mann-Whitney

B1. Gộp hai mẫu và sấp xếp $n_1 + n_2$ phần tử theo thứ tự tăng dần.
B2. Tìm hạng của x_i và y_i. Tìm R_1 và R_2 là tổng hạng của mẫu 1 và 2.
B3.

$$Z_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1;$$

$$Z_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2;$$

Gọi $Z = \min(Z_1, Z_2)$

$$EZ = \frac{n_1 n_2}{2}; DZ = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

$$z = \frac{Z - EZ}{\sqrt{DZ}}$$

Miền Tiêu chuẩn:

$$S = \{|z| \geq z(\frac{\alpha}{2})\}$$

Tiêu chuẩn Wilconxon

B1. Tìm hiệu $d_i = x_i - y_i, i = 1, 2, ..., n$ bỏ những giá trị bằng không. Gọi n^+ là số các số hạng $d_i \neq 0$
B2. Tìm hạng của $|d_i|$. B3. Tìm $T^- = \sum_{i:d_i<0} rank(|d_i|); T^+ = \sum_{i:d_i>0} rank(|d_i|)$
Gọi $T = \min(T^-, T^+)$

$$ET = \frac{n^+(n^+ + 1)}{4}; DT = \frac{n^+(n^+ + 1)(2n^+ + 1)}{24}$$

$$z = \frac{T - ET}{\sqrt{DT}}$$

Miền Tiêu chuẩn:

$$S = \{|z| \geq z(\frac{\alpha}{2})\}$$

Công thức

$$z = \frac{p_1^* - p_2^*}{\sqrt{p^*(1-p^*)(\frac{1}{n_1} + \frac{1}{n_2})}};$$

$$S_1 = \{|z| \geq z(\frac{\alpha}{2})\} \quad S_2 = \{z \geq z(\alpha)\} \quad S_3 = \{z \leq -z(\alpha)\}$$

Ví dụ 18

Để thăm dò ý kiến của nhân dân về một điều khoản nào đó, người ta chon ra hai mẫu đại diện ở thành thị và nông thôn.
Ở thành thị $n_1 = 500; m_1 = 320$ ý kiến ủng hộ.
Ở nông thôn $n_2 = 400; m_2 = 300$ ý kiến ủng hộ.
Với mức ý nghĩa $\alpha = 0.05$ có thể kết luận người dân ở nông thôn ủng hộ điều khoản này cao hơn thành thị hay không?

VI. Tiêu chuẩn phù hợp χ^2

Công thức

$$\chi^2 = \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i} = \frac{1}{n}\sum_{i=1}^{k} \frac{m_i^2}{p_i} - n; \quad S = \{\chi^2 \geq \chi^2_{k-1}(\alpha)\}$$

Ví dụ 19

Theo thống kê thì trước nghị định 36CP tỷ lệ các vụ tai nạn giao thông đường bộ do người đi bộ, xe đạp, xe máy, ô tô gây ra ở thành phố Z tương ứng là $10\%, 15\%, 60\%, 15\%$. Sau 3 tháng thực hiện nghi định này, Ở thành phố Z đã xảy ra 250 vụ tai nạn giao thông đường bộ, trong đó có 40 vụ do lỗi người đi bộ, 60 vụ do lỗi người đi xe đạp, 120 vụ do lỗi người đi xe máy và 30 vụ do lỗi người đi ô tô gây ra.

a. Với mức ý nghĩa $\alpha = 0.05$ có thể kết nói rằng sau nghi định 36CP nguyên nhân gây ra tai nạn giao thông đường bộ đã thay đổi so với trước hay không? Rút ra ý nghĩa thực tiễn gì từ kết luận nhận được.

b. Với mức ý nghĩa $\alpha = 0.05$ có thể kết luận rằng tỷ lệ các vụ tai nạn do người đi bộ và đi xe đạp tăng lên so với trước còn tỷ lệ tai nạn do người đi xe náy và ô tô giảm đi so với trước hay không?

Áp dụng tiêu chuẩn phù hợp χ^2 để kiểm tra sự phù hợp của phân bố

Chia miền giá trị mẫu thành các phần $S_1, ..., S_k$ sau đó xác định $p_1 = P(X \in A_1), ..., p_k = P(X \in A_k)$ và áp dụng tiêu chuẩn χ^2 đe kiểm định.

Ví dụ 20 (VD 12 Tr188)

Tiến hành đo chiều cao của 100 cây bạch đàn trong một khu rừng trồng bạch đàn của một lâm trường ta thu được kết quả sau:

Khoảng chiều cao (m)	số cây	Khoảng chiều cao (m)	số cây
8.275 – 8.325	1	8.625 – 8.675	17
8.325 – 8.375	2	8.675 – 8.725	12
8.375 – 8.425	4	8.725 – 8.775	9
8.425 – 8.475	5	8.775 – 8.825	7
8.475 – 8.525	8	8.825 – 8.875	6
8.525 – 8.575	10	8.875 – 8.925	0
8.575 – 8.625	18	8.925 – 8.975	1

Hãy kiểm tra giả thiết cho rằng chiều cao của các cây bạch đàn là tuân theo phân bố chuẩn ($\alpha = 0.05$)?

Bài làm

Khoảng $S_i(a_i, a_{i+1})$	m_i	p_i	$\frac{(m_i - np_i)^2}{np_i}$
$(-\infty, 8.425)$	7	0.0548	0.4216
$(8.425, 8.475)$	5	0.0583	0.1182
$(8.475, 8.525)$	8	0.0930	0.1817
$(8.525, 8.575)$	10	0.1295	0.6720
$(8.575, 8.625)$	18	0.1484	0.6729
$(8.625, 8.675)$	17	0.1528	0.1936
$(8.675, 8.725)$	12	0.1735	0.1365
$(8.725, 8.775)$	9	0.1004	0.1077
$(8.775, 8.825)$	7	0.0650	0.0385
$(8.825, +\infty)$	7	0.0643	0.0505
$\sum$	100		$\chi^2 = 2.5923$

Với $\alpha = 0.05$ thì $\chi^2_7(0.05) = 14.0671$, vậy $\chi^2 < \chi^2_7(0.05)$ nên ta chấp nhận giả thiết, tức biến ngẫu nhiên chiều cao của cây bạch đàn là tuân theo phân bố chuẩn với độ tin cậy là 95%.

Áp dụng tiêu chuẩn phù hợp χ^2 để kiểm tra sự phù hợp của phân bố

Chú ý

Trong trường hợp các phân bố chưa cho một số tham số ta cần thay thế các tham số đó bằng các ước lượng của nó dựa trên mẫu đã cho. Chẳng hạn
Nếu $X \sim N(\mu, \sigma^2)$ thay μ bằng $\overline{X}$ và σ^2 bằng s^2 hoặc σ_X^2.
Nếu $X \sim P(\lambda)$ thay λ bằng $\overline{X}$.
Nếu $X \sim B(n, p)$ thay thế p bằng $\overline{X}$ hoặc m/n.
Nếu $X \sim E(\lambda)$ thay λ bằng $\frac{1}{\overline{X}}$.
Nếu $X \sim U[a, b]$ thay $a = \overline{X} - \sqrt{3}s, b = \overline{X} + \sqrt{3}s$
Khi đó MTC:

$$S = \{\chi^2 \geq \chi^2_{k-r-1}(\alpha)\}$$

Kiểm tra tính độc lập

Công thức

Cộng tổng hàng: $hgi = \sum_{j=1}^{s} n_{ij}$, tổng cột: $cotj = \sum_{i=1}^{r} n_{ij}$

Tính $\chi^2 = n(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{hgicotj} - 1)$

Miềm tiêu chuẩn $S = \{\chi^2 \geq \chi^2_{(r-1)(s-1)}(\alpha)\}$

Ví dụ

Nghiên cứu về sự phụ thuộc giữa đạo đức của trẻ vị thành niên và hoàn cảnh gia đình của họ, ta có một mẫu được điều tra ngẫu nhiên như sau

	Bố hoặc mẹ đã mất	Bố mẹ ly hôn	còn cả bố, mẹ
Ngoan	20	15	50
Hư	8	12	10
Phạm tội	6	9	5

Với mức ý nghĩa $\alpha = 0.1$ có thể kết luận tình trạng đạo đức của trẻ ở tuổi vị thành niên phụ thộc vào hoàn cảnh gia đình hay không?

So sánh nhiều tỷ lệ

I. Công thức

Cộng tổng hàng: $hgi = \sum_{j=1}^{2} n_{ij}$, tổng cột: $cotj = \sum_{i=1}^{k} n_{ij}$

Tính $\chi^2 = n(\sum_{i=1}^{2}\sum_{j=1}^{k} \frac{n_{ij}^2}{hg_i cot_j} - 1)$

Miềm tiêu chuẩn $S = \{\chi^2 \geq \chi^2_{k-1}(\alpha)\}$

II. Ví dụ

Một hãng sản suất ô tô muốn tìm hiểu xem có sự phụ thuộc nào giữa giới tính của người sở hữu và kiểu dáng xe ô tô hay không. Một mẫu ngẫu nhiên gồm 2000 chủ sơ hữu ô tô được chon và phân loại như sau

	I	II	III
Nam	350	270	380
Nữ	340	400	260

Với mức ý nghĩa $\alpha = 0.025$, tỷ lệ nữ dùng 3 loại xe trên có như nhau hay không?

I. Công thức

Từ số liệu của mẫu chung của X, Y, tìm phân bố mẫu của từng thành phần. (Cộng tổng hàng: $hgi = \sum_{j=1}^{s} n_{ij}$, tổng cột: $cotj = \sum_{i=1}^{r} n_{ij}$)

Tính $\overline{X}, \overline{Y}, \sigma_X, \sigma_Y$ và $\overline{XY}$ trong đó $\overline{XY} = 1/n. \sum_{i=1}^{s} \sum_{j=1}^{n} m_{ij} x_i y_j$

Khi đó hệ số tương quan mẫu $r = \dfrac{\overline{XY} - \overline{XY}}{\sigma_X \sigma_Y}$

II. Ví dụ

Điều tra số trẻ em thất học (X) và số trẻ em hư (Y) ở 30 địa phương ta có

X Y	16	18	20	22	24	26
80	3	1				
90		4	3	1		
100		1	3	1		
110			3	1	1	
120				2	2	1
130					2	1

Tiếp Ví dụ

a. Tìm hệ số tương quan mẫu r?
b. Có nhận xét gì về sự phụ thuộc giữa số trẻ em hư và trẻ em thất học qua kết quả r vừa tính?

I. Công thức đường hồi quy

Đường hồi quy của y theo x: $y - \overline{Y} = r.\frac{\sigma_y}{\sigma_x}(x - \overline{X})$ với sai số bình phương trung bình là $\sigma^2_{y/x} = \sigma^2_y(1 - r^2)$

Đường hồi quy của x theo y: $x - \overline{X} = r.\frac{\sigma_x}{\sigma_y}(y - \overline{Y})$ với sai số bình phương trung bình là $\sigma^2_{x/y} = \sigma^2_x(1 - r^2)$

II. Ví dụ

c. Xây dựng đường hồi quy tuyến tính TN của trẻ em hư theo số trẻ em thất học?
d. Nếu biết ở một địa phương có số trẻ em thất học là $x = 105$ hãy dự đoán số trẻ em hư của địa phương?

Cách tìm r và hệ số A, B của đường hồi quy bằng máy tính Fx 500,570 MS

Mở chương trình: Mode $\rightarrow$ 3$\rightarrow$ 1. (Fx500). Mode $\rightarrow$ Mode $\rightarrow$ 2 $\rightarrow$ 1. (Fx570)
Nhập số liệu: $x_1 \rightarrow, \rightarrow y_1 \rightarrow$ *shift* $\rightarrow; \rightarrow m_1 \rightarrow DT$
.....
$x_k \rightarrow, \rightarrow y_k \rightarrow$ *shift* $\rightarrow; \rightarrow m_k \rightarrow DT$
Kết quả: *shift* $\rightarrow 2 \rightarrow \triangleright\triangleright \rightarrow 1 \rightarrow =$ (cho A)
shift $\rightarrow 2 \rightarrow \triangleright\triangleright \rightarrow 2 \rightarrow =$ (cho B)
shift $\rightarrow 2 \rightarrow \triangleright\triangleright \rightarrow 3 \rightarrow =$ (cho r)

Cách tìm r và hệ số A, B của đường hồi quy bằng máy tính Fx 570 ES

Mở chương trình: Shift$\rightarrow$ Mode $\rightarrow$ Mode $\rightarrow \triangleright 4 \rightarrow 1$
Mode $\rightarrow 3 \rightarrow 2$.
Nhập số liệu: Nhập x_i, y_i, m_i. Kết quả:
shift $\rightarrow 1 \rightarrow 7(5) \rightarrow 1 \rightarrow =$ (cho A)
shift $\rightarrow 1 \rightarrow 7(5) \rightarrow 2 \rightarrow =$ (cho B)
shift $\rightarrow 1 \rightarrow 7(5) \rightarrow 3 \rightarrow =$ (cho r)

II. Ví dụ