

BÀN VỀ TIÊU CHUẨN : TCVN 9153:2012

“CÔNG TRÌNH THỦY LỢI - PHƯƠNG PHÁP CHÍNH LÝ KẾT QUẢ THÍ NGHIỆM MẪU ĐẤT” [1]

Huyền Kim Ân

Công ty CP TVXD điện 4 (PECC 4)

MỞ ĐẦU: Tiêu chuẩn TCVN 9153:2012 “Công trình thủy lợi – phương pháp chính lý kết quả thí nghiệm mẫu đất” được ban hành theo quyết định số 3576/QĐ-BKHCHN ngày 27/12/2012 là một tiêu chuẩn quốc gia đã được áp dụng đến nay gần 3 năm (nếu nói xa hơn nữa là tiêu chuẩn này được soạn lại từ TCXD 74:1987 ban hành năm 2002, tức đã gần 13 năm); trong quá trình áp dụng thấy rằng còn nhiều điều cần bàn luận thêm. Những sai sót trong tiêu chuẩn đã được các chuyên gia chuyên ngành có nhiều đóng góp ý kiến; phạm vi bài viết này, đề cập một vấn đề mà tiêu chuẩn chưa nêu ra, chưa có hướng dẫn nhưng trong thực tế luôn gặp phải.

Vấn đề xuất phát từ mục 3.3 – Quy định chung của [1] ghi: “Phương pháp chính lý thống kê sử dụng trong tiêu chuẩn tuân theo luật phân phối chuẩn”.

Câu văn trong dấu “... “ trên là không đúng với lý thuyết xác suất thống kê vì, Phương pháp ... mà tuân theo luật phân phối chuẩn có nghĩa gì ?. Đây, có lẽ là do dịch thuật !

Ý của câu văn trên phải được hiểu: Các biến thu thập trong quá trình khảo sát đưa vào chính lý thống kê (sau khi xử lý số liệu thô) phải tuân theo luật phân bố chuẩn (Gaussian distribution); đây là một giả định thống kê quan trọng. Nhưng, nếu dữ liệu không tuân theo luật phân bố chuẩn thì giải quyết như thế nào ? Trong tiêu chuẩn này chưa có câu trả lời !

Lấy một ví dụ: Tại “mục 4.1.6 – Các bước chính lý. Kiểm tra sự cần thiết phải phân chia tiếp đơn nguyên địa chất công trình bằng tiêu chuẩn “ t ” theo công thức 6” là một ví dụ (chú ý công thức 6 trong tiêu chuẩn viết sai-bạn đọc có thể kiểm tra lý thuyết xác suất thống kê).

Thực chất, đây là bài toán kiểm định t (còn gọi “t.test”) nhằm để trả lời câu hỏi: có hay không có sự khác nhau giữa 2 số trung bình của 2 mẫu? mục đích là để phân chia đơn nguyên địa chất. Phương pháp kiểm định t này dựa vào một số giả định thống kê sau:

- Hai nhóm so sánh phải hoàn toàn độc lập nhau;
- Biến so sánh phải tuân theo luật phân phối chuẩn (Gaussian distribution);
- Phương sai 2 nhóm bằng nhau (hoặc tương đương);
- Biến phải được chọn ngẫu nhiên (random sample);

Trường hợp mẫu không thỏa mãn 1 trong những giả định trên thì tiêu chuẩn TCVN 9153:2012 chưa đưa ra hướng dẫn thực hiện. Đây là cái thiếu của tiêu chuẩn mà nội dung bài viết này đề xuất cách giải quyết.

(Thông thường, nếu dữ liệu không tuân luật phân phối chuẩn hoặc phương sai không tương đương thì có thể hoán chuyển dữ liệu sang logaric hoặc hàm số mũ, mục tiêu là làm cho thông số mới sẽ tuân luật phân phối chuẩn; sau đó tính toán các bước như tiêu chuẩn trên hướng dẫn-Lý thuyết xác suất thống kê).

Tuy nhiên, cũng có những trường hợp dữ liệu không thể hoán chuyển sang bất kỳ dạng nào để thông số mới tuân luật phân phối chuẩn được; trường hợp này sử dụng phương pháp *Bootstrap*.

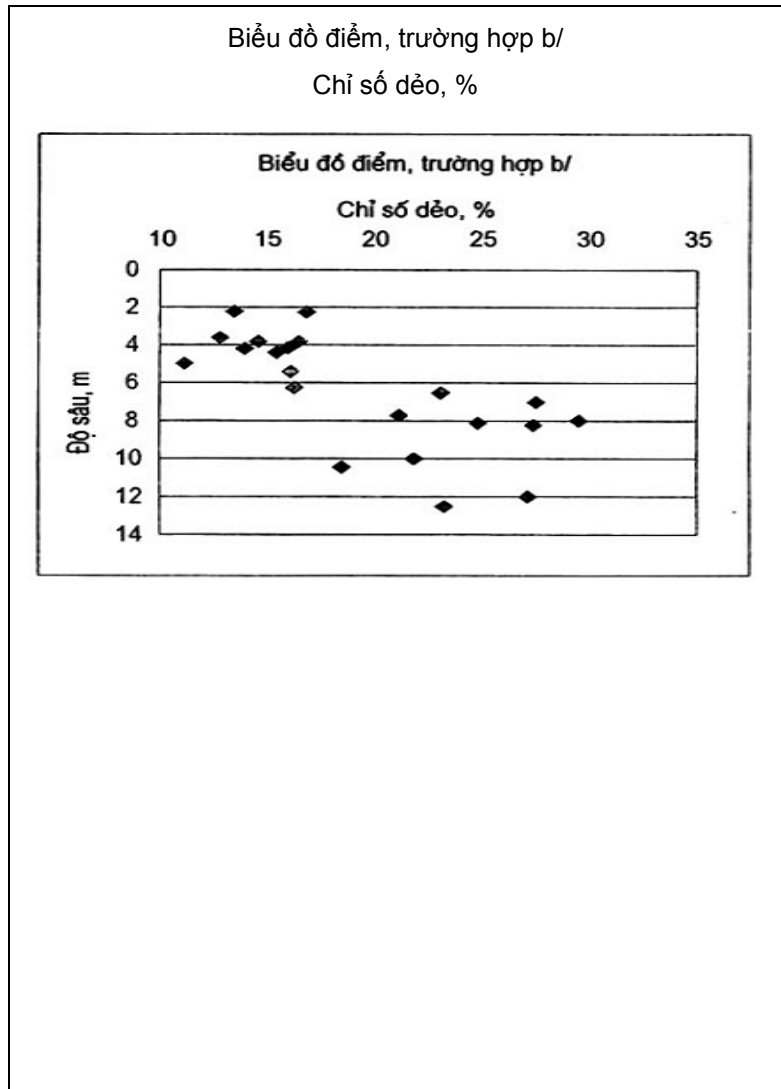
GIỚI THIỆU PHƯƠNG PHÁP.

Để giới thiệu phương pháp; trong thực hành sẽ sử dụng dữ liệu từ bảng B.2, phụ lục B, TCVN 9153:2012; các bước tính toán kiểm định t.test được sử dụng theo hướng dẫn “phân tích dữ liệu thống kê với R” [2]; không theo trình tự hướng dẫn của tiêu chuẩn với mục đích giảm khối lượng tính toán (thực chất kết quả là như nhau).

Chỉ số dẻo cho bảng sau:

B.2 Trường hợp hai lớp đất.

Hố khoan	Độ sâu, m	Chỉ số dẻo, %
K1	4	16
K1	6,8	27,6
K2	3,8	14,7
K2	6,3	23,1
K2	11,5	27,1
K3	4	14,1
K3	6	16,3
K3	8	27,4
K4	3,8	16,5
K4	7,8	24,9
K5	3,5	12,9
K5	9,5	21,9
K5	12	23,2
K5	7,8	29,5
K6	2,2	13,5
K6	5,2	16,1
K7	4,2	15,6
K7	2,2	16,9
K7	7,5	21,2
K8	4,8	11,2
K8	10	18,4



Câu hỏi đặt ra là: Có hay không có sự khác biệt chỉ số dẻo của 2 lớp đất kề nhau ? mục đích là hợp nhất hay chia thành 2 đơn nguyên địa chất riêng biệt?.

Giới thiệu 2 cách trả lời câu hỏi trên: Như hướng dẫn ở mục 4.1.6 TCVN 9153: 2012 (kiểm định t) và phương pháp Bootstrap.

1. Kiểm định t-test (Trình tự các bước thực hiện theo ngôn ngữ R, thuận lợi và nhanh chóng hơn là cách tính tay và tra bảng như tiêu chuẩn trên hướng dẫn)
Gọi B_1 chỉ số dẻo ở độ sâu từ 0 – 6 m;

B₂ là chỉ số dẻo ở độ sâu từ 6.5 – 14 m. Từ bảng B.2 nhập vào R như sau:

B₁ = (16,14.7,23.1,14.1,16.3,16.5,12.9,13.5,16.1,15.6,16.9,11.2)

B₂ = (27.6,27.1,27.4,24.9,21.9,23.2,29.5,21.2,18.4)

Kiểm định t với ngôn ngữ R chỉ bằng lệnh đơn giản là: t.test (B1,B2), ta có kết quả như sau:

Output:

```
t.test(B1,B2)

welch Two Sample t-test

data: B1 and B2
t = -6.0855, df = 15.014, p-value = 2.077e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.155764 -5.849791
sample estimates:
mean of x mean of y
15.57500 24.57778
```

Từ Output thấy:

t = - 6.08 (nằm ngoài khoảng giá trị “t- lý thuyết”; bậc tự do df = 15, p = 2.07*10⁻⁵ => sự khác biệt giữa 2 lớp có ý nghĩa thống kê (nghĩa là có sự khác biệt giữa B₁ và B₂);

Khoảng tin cậy 95% của khác biệt B₁ với B₂ từ (- 12.15) đến (- 5.85);

Giá trị trung bình của B₁: 15.57, của B₂ = 24.57 (độ khác biệt của trung bình là 9.0)

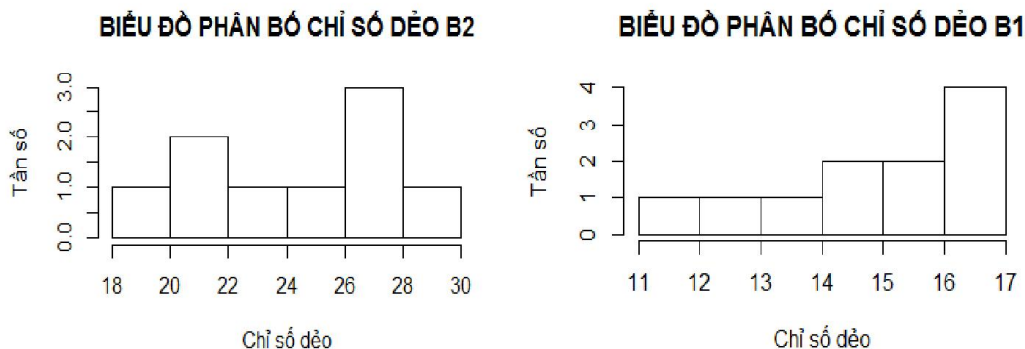
Tuy nhiên, kết luận trên có thể sai nếu dữ liệu (B₁, B₂) không đáp ứng giả định thống kê !

Kiểm tra các giả định thống kê:

+ Phương sai 2 biến phải bằng nhau hoặc tương đương:

Ở đây, Phương sai của: var(B₁) = 8.5238 và var(B₂) = 13.304 => sai khác không lớn.

+ Về luật phân bố chuẩn: Vẽ biểu đồ phân bố 2 biến B₁, B₂, trong R chỉ đơn giản bằng lệnh “hist(B1)” như hình sau:



Với biểu đồ phân bố như trên, kết luận rằng dữ liệu không tuân luật phân phối chuẩn (nếu dữ liệu tuân theo luật phân phối chuẩn thì biểu đồ tần suất có dạng hình chuông).

Như vậy, kiểm định t vấp phải vấn đề giả định thống kê !

2. Phương pháp Bootstrap [3].

Phương pháp phân tích Bootstrap dựa vào cách tái chọn mẫu và mô phỏng dựa vào dữ liệu thực tế để tìm phân bố cho 1 chỉ số thống kê nào đó. Các bước như sau:

- Lấy mẫu nhóm 1 (B_1), nhóm 2 (B_2);
 - Tính hiệu số của 2 số trung bình;
 - Lặp lại các bước trên B lần (có thể đến 10.000 lần);
 - Xét phân bố để xác định khoảng tin cậy 95%.
- Cú pháp lệnh trong "R" và output như sau:

```
> B1 = c(16,14.7,23.1,14.1,16.3,16.5,12.9,13.5,16.1,15.6,16.9,11.2)
> B2 = c(27.6,27.1,27.4,24.9,21.9,23.2,29.5,21.2,18.4)
> n1=length(B1)
> n2=length(B2)
> B=10000
> difference=numeric(B)
> for (i in 1:B){
+ bs.B1<-sample(B1,n1,replace=TRUE)
+ bs.B2<-sample(B2,n2,replace=TRUE)
+ difference[i]=mean(bs.B2)-mean(bs.B1)
+ }
> quantile(difference,probs=c(0.025,0.5,0.975))
      2.5%      50%      97.5%
6.197153  9.027778 11.627847
> # Phân bố của khác biệt của 2 nhóm
> hist(difference,breaks=20,col="blue",border="white")
```

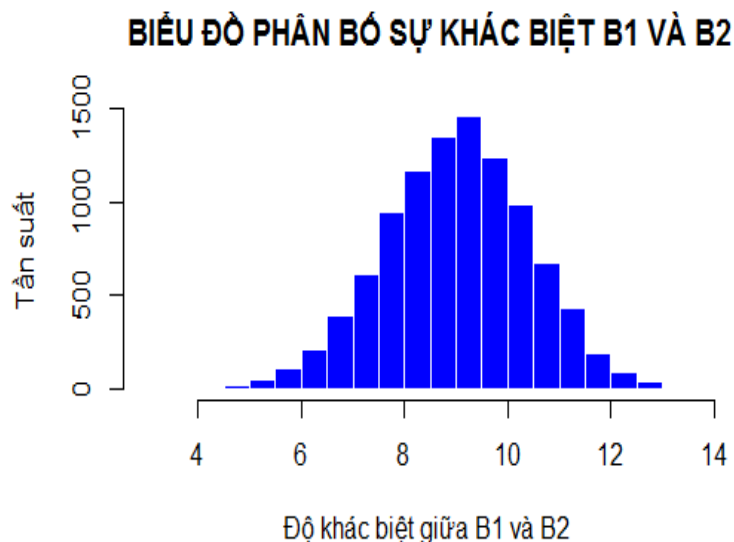
Output trên cho thấy:

Khác biệt chỉ số đều trung bình của 2 nhóm là 9.02;

Khoảng tin cậy 95% khác biệt giữa B_1 và B_2 từ 6.197 đến 11.627 (Xem phân bố như hình sau); ***từ kết quả này, có cơ sở để khẳng định rằng có sự khác biệt giữa 2 nhóm.***

Ở kết quả "t-test" trên, độ khác biệt này dao động từ: 5.849 đến 12.155 . Trường hợp này, cả 2 phương pháp có kết luận tương đương; tuy nhiên như đã phân tích trên, t.test không đáp ứng giả định thống kê.

Phân bố của khác biệt giữa 2 nhóm như hình sau:



Dựa vào sự khác biệt về chỉ số dẻo của 2 nhóm như kết quả kiểm định trên là cơ sở phân chia địa tầng thành 2 đơn nguyên.

KẾT LUẬN:

Kiểm định t đơn giản trong tính toán, không cần máy tính bằng cách tính tay theo trình tự như hướng dẫn của tiêu chuẩn; tuy nhiên dễ mắc phải sai lầm do không chú ý những giả định thống kê đằng sau của phương pháp.

Hơn nữa, nếu dữ liệu thu thập được không đáp ứng các giả định, tiêu chuẩn trên cũng chưa đưa ra hướng dẫn để xử lý.

Đề xuất phương pháp Bootstrap thay cho phương pháp kiểm định t (t.test), nếu bằng máy tính với bất kỳ phần mềm nào về xác suất thống kê như SAS, SPSS... đều nhanh chóng; lợi thế của phương pháp là không bị ràng buộc bởi những giả định thống kê./.

Vài lời cảm ơn: Trong quá trình chuẩn bị dự thảo bài viết trên, tác giả xin trân trọng cảm ơn Cô: Tôn Nữ Mỹ Liên – Phó GD Trung tâm thí nghiệm chuyên ngành xây dựng - PECC 4, đã có những góp ý quý báu giúp cho bản thảo được hoàn chỉnh hơn.

Tài liệu tham khảo:

[1]. TCVN 9153:2012 “ Công trình thủy lợi – Phương pháp chỉnh lý kết quả thí nghiệm mẫu đất”.

[2]. GS Nguyễn Văn Tuấn-“Phân tích dữ liệu với R” NXB tổng hợp TP HCM, 10/2014.

[3] Huỳnh Kim Ân “Giới thiệu phương pháp mô phỏng từ dữ liệu thực tế để suy luận cho quần thể” vncold.vn [04/06/15].