BÀI TẬP VÀ THỰC HÀNH

Bài 3.1: Hãy dùng lệnh FREQ để đưa ra phân bố tần số của những bệnh nhân xuất huyết và không xuất huyết.

Bài 3.2: Hãy tính tỷ lệ các bệnh nhân có mức độ BC là thấp, trung bình, cao.

Bài 3.3: Dùng lệnh TABLES để đưa ra bảng phân bố tần số, tỷ lệ của bệnh nhân xuất huyết và không xuất huyết với giới tính. Hãy đọc kết quả tìm được và rút ra kết luận về thống kê như thế nào?

Bài 3.4: Dùng STATCAL để giải bài toán sau:

Điều tra tình hình mắc ba bệnh (B) B1, B2, B3 tại hai phân xưởng (FX) I và II của nhà máy X thu được kết quả sau:

FX \ B	B_1	B_2	B_3	m_{i0}
I	588	369	89	1046
II	304	171	50	525
m_{0j}	892	540	139	1571

Tỷ lệ ba bệnh tại hai phân xưởng có như nhau không?

Hãy đọc kết quả tìm được và rút ra kết luận về thống kê như thế nào?

Bài 4

SO SÁNH TRUNG BÌNH VÀ TÍNH TƯƠNG QUAN TUYẾN TÍNH TRONG EPI–INFO 6.04

MỤC TIÊU

1. *Thực hiện được lệnh Means để tính các tham số đặc trưng thực nghiệm cho một biến định lượng*
2. *Thực hiện được lệnh Means để so sánh hiệu quả trước sau (so sánh từng cặp).*
3. *Thực hiện được lệnh Means để so sánh trung bình của nhiều nhóm nghiên cứu.*
4. *Thực hiện được lệnh REGRESS để tính tương quan hồi quy tuyến tính (Linear regression).*

1. LỆNH MEANS – TÍNH TRUNG BÌNH VÀ SO SÁNH TRUNG BÌNH CỦA CÁC NHÓM NGHIÊN CỨU

1.1. Dùng lệnh Means để tính các tham số đặc trưng thực nghiệm cho một biến định lượng

Trong khi lệnh TABLES dùng để so sánh các tỷ lệ hoặc kiểm định tính độc lập của hai đặc tính về chất, thì lệnh MEANS thực hiện các thuật toán với các biến định lượng (biến định lượng là biến mà các giá trị của nó là số liên tục). Ví dụ: chiều cao, cân nặng, tuổi, hồng cầu, bạch cầu, v.v…

Cú pháp: MEANS <tên biến>

Kết quả của lệnh MEANS với một biến định lượng sẽ hiển thị lên màn hình gồm:

- Bảng phân phối tần số (một chiều).
- Tổng giá trị của biến (Sum).
- Trị số trung bình (Mean).
- Phương sai (Variance), Độ lệch chuẩn (Std Dev), Sai số chuẩn (Std Err).
- Giá trị nhỏ nhất (Minimum), Giá trị tại điểm 25% (25%ile), Trung vị (Median), Giá trị tại điểm 75% (75%ile), Giá trị lớn nhất (Maximum), Giá trị hay gặp nhất (Mode).

Ví dụ: MEANS BC ↵

Máy sẽ hiện kết quả:

BC	Freq	Percent	Cum.
0.6	1	1.2%	1.2%
1.4	1	1.2%	2.3%
2.3	1	1.2%	3.5%
....	...	....	...
43.0	1	1.2%	97.7%
44.5	1	1.2%	98.8%
45.7	1	1.2%	100.0%
Total	86	100.0%	

Total	Sum	Mean	Variance	Std Dev	Std Err
86	1408	16.372	108.036	10.394	1.121

Minimum	25%ile	Median	75%ile	Maximum	Mode
0.600	10.100	13.250	20.000	45.700	12.700

Student's "t", testing whether mean differs from zero.
T statistic = 14.607, df = 85 p-value = 0.00000

Ta có:

n	= 86	Giá trị nhỏ nhất	= 0.6
BC trung bình	= 16.372	Giá trị lớn nhất	= 45.7
Phương sai	= 108.036	Trung vị	= 13.25
Độ lệch	= 10.394	Giá trị hay gặp nhất	= 12.7

Test "t" ở đây kiểm định xem giá trị trung bình của biến đang xét khác biệt có ý nghĩa thống kê so với giá trị "0" không ?

- Nếu trị số của p (p–value) ≥ 0.05 thì ta kết luận là "Khác biệt không có ý nghĩa thống kê với độ tin cậy 95%".
- Nếu trị số của p (p–value) < 0.05 thì ta kết luận là "Khác biệt có ý nghĩa thống kê với độ tin cậy 95%".

Chú ý: Nếu sau lệnh FREQ ta đưa vào tên của biến mà các giá trị của nó là số thì máy cũng đưa ra kết quả giống như khi sử dụng lệnh MEANS.

Ví dụ: FREQ BC ↵ ta cũng có kết quả như trên.

1.2. Dùng lệnh Means để so sánh hiệu quả trước sau (so sánh từng cặp)

Trong bài toán so sánh hiệu quả trước–sau, áp dụng lệnh MEANS cho biến HIEU, Test "t" sẽ kiểm định xem giá trị trung bình của biến HIEU khác biệt có ý nghĩa thống kê so với giá trị "0" hay không.

Ví dụ: Định lượng Protein toàn phần trong huyết thanh bệnh nhi suy dinh dưỡng (đơn vị g/l) trước điều trị (TRDT) và sau điều trị (SAUDT), thu được số liệu sau:

TRDT 55.8 53.3 30.1 51.0 37.8 68.6 57.7 59.1 49.4 35.4 53.4 42.7 21.2 28.3 57.3 42.4 61.4

SAUDT 60.4 58.7 28.9 48.0 39.7 68.8 57.5 70.4 56.8 40.6 57.3 44.3 32.2 47.7 77.0 55.1 66.1

Hỏi: Phương pháp điều trị có hiệu quả không?

Để đánh giá phương pháp điều trị có hiệu quả không, ta phải áp dụng thuật toán so sánh từng cặp.

Cách làm: Nhập lượng Protein toàn phần trước điều trị vào biến TRDT.

Nhập lượng Protein toàn phần sau điều trị vào biến SAUDT.

Sau đó thực hiện lệnh: DEFINE HIEU ###.# ↵

LET HIEU=SAUDT – TRDT ↵

MEANS HIEU ↵

Có kết quả sau:

HIEU	Freq	Percent	Cum.
-3.0	1	5.9%	5.9%
-1.2	1	5.9%	11.8%
-0.2	1	5.9%	17.6%
0.2	1	5.9%	23.5%
1.6	1	5.9%	29.4%
1.9	1	5.9%	35.3%
3.9	1	5.9%	41.2%
4.6	1	5.9%	47.1%
4.7	1	5.9%	52.9%
5.2	1	5.9%	58.8%
5.4	1	5.9%	64.7%
7.4	1	5.9%	70.6%
11.0	1	5.9%	76.5%
11.3	1	5.9%	82.4%
12.7	1	5.9%	88.2%
19.4	1	5.9%	94.1%
19.7	1	5.9%	100.0%
Total	17	100.0%	

Total	Sum	Mean	Variance	Std Dev	Std Err
17	105	6.153	44.809	6.694	1.624

Minimum	25%ile	Median	75%ile	Maximum	Mode
-3.000	1.600	4.700	11.000	19.700	-3.000

Student's "t", testing whether mean differs from zero.
T statistic = 3.790, df = 16, p-value = 0.00161

Sau điều trị, lượng Protein tăng trung bình là 6.153 g/l và lượng Protein tăng trung bình này là thực sự có ý nghĩa thống kê với độ tin cậy 99% vì giá trị p=0.00161. Vậy phương pháp điều trị là có hiệu quả.

Dùng lệnh Means để so sánh trung bình của nhiều nhóm nghiên cứu.

Lệnh MEANS còn khảo sát sự tương quan giữa một đặc tính định lượng và một đặc tính định tính. Mỗi giá trị của biến định tính là một tiêu chuẩn phân nhóm các giá trị định lượng để so sánh. Nếu biến định tính có 2 giá trị khác nhau thì ta có kết quả của thuật toán so sánh hai trung bình. Nếu biến định tính có từ 3 giá trị khác

nhau trở lên thì ta có kết quả của thuật toán so sánh nhiều trung bình.

Cú pháp: MEANS <tên biến định lượng> <tên biến định tính>

Ví dụ: So sánh lượng SGPT trung bình giữa hai nhóm bệnh nhân.

Nhóm 1 – Hôn mê gan do bệnh cấp tính.

2 – Hôn mê gan do bệnh mạn tính.

Lệnh: SELECT SGPT>0 ↵
MEANS SGPT NHOM ↵

Kết quả:

NHOM	Obs	Total	Mean	Variance	Std Dev
1	57	4071	71.421	1877.212	43.327
2	15	499	33.267	496.210	22.276
Difference			38.154		

NHOM	Minimum	25%ile	Median	75%ile	Maximum	Mode
1	12.000	35.000	72.000	102.000	210.000	20.000
2	15.000	18.000	26.000	38.000	91.000	21.000

ANOVA
(For normally distributed data only)

Variation	SS	df	MS	F statistic	p-value	t-value
Between	17287.116	1	17287.116	10.798	0.001962	3.285973
Within	112070.828	70	1601.012			
Total	129357.944	71				

```
            Bartlett's test for homogeneity of variance
  Bartlett's chi square = 7.304  deg freedom = 1  p-value = 0.006879
    Bartlett's Test shows the variances in the samples to differ.
       Use non-parametric results below rather than ANOVA.
Mann-Whitney or Wilcoxon Two-Sample Test(Kruskal-Wallis
                    test for two groups)
Kruskal-Wallis H (equivalent to Chi square) =      11.734
                        Degrees of freedom =           1
                                   p value =    0.000614
```

Lần đầu tiên nhìn kết quả trên chắc chắn ta sẽ cho là nhiều con số quá. Nhưng đối với ta chỉ cần quan tâm đến một vài giá trị cần thiết như:

- Kích thước n của từng nhóm.
- Giá trị trung bình, phương sai và độ lệch của từng nhóm.
- Trị số p của test để so sánh phương sai ở các nhóm – p–value của Bartlett's Test.
- Trị số p của test để so sánh trung bình ở các nhóm – p–value của ANOVA Test hoặc Kruskal–Wallis H test.

Ta có quy tắc:

- Trong trường hợp trị số p của Bartlett's Test ≥ 0.05, điều đó có nghĩa là phương sai của các nhóm khác biệt nhau không có ý nghĩa thống kê. Để so sánh các trung bình, ta phải xem tiếp trị số p của

ANOVA test, nếu trị số p của ANOVA test ≥ 0.05 thì kết luận "giá trị trung bình của các nhóm khác biệt không có ý nghĩa thống kê với độ tin cậy 95%", còn nếu trị số p của ANOVA test < 0.05 thì kết luận "giá trị trung bình của các nhóm khác biệt có ý nghĩa thống kê với độ tin cậy 95%".

– Trong trường hợp trị số p của Bartlett's Test < 0.05, điều đó có nghĩa là phương sai của các nhóm khác biệt nhau có ý nghĩa thống kê. Để so sánh các trung bình, ta phải xem tiếp trị số p của Kruskal–Wallis H test, nếu trị số p của Kruskal–Wallis H test ≥ 0.05 thì kết luận "giá trị trung bình của các nhóm khác biệt không có ý nghĩa thống kê với độ tin cậy 95%", còn nếu trị số p của Kruskal–Wallis H test < 0.05 thì kết luận "giá trị trung bình của các nhóm khác biệt có ý nghĩa thống kê với độ tin cậy 95%".

Ở ví dụ trên ta thấy:

Lượng SGPT

Nhóm	n	trung bình	phương sai	độ lệch chuẩn
Hôn mê gan do bệnh cấp tính	57	71.421	1877.212	43.327
Hôn mê gan do bệnh mạn tính	15	33.267	496.210	22.276
Trị số p của Bartlett's test:		p–value = 0.006879		
Trị số p của Kruskal–Wallis H test:		p–value = 0.000614		

Phương sai của hai nhóm khác biệt nhau có ý nghĩa thống kê với độ tin cậy 99% vì p–value của Bartlett's Test < 0.01. Phương sai của nhóm hôn mê gan do bệnh cấp tính lớn hơn phương sai của nhóm hôn mê gan do bệnh mạn tính. Trong nhóm hôn mê gan do bệnh cấp tính, các giá trị về lượng SGPT có phân bố tản mạn hơn so với lượng SGPT của nhóm hôn mê gan do bệnh mạn tính.

Lượng SGPT trung bình giữa hai nhóm hôn mê gan do bệnh cấp tính và do bệnh mạn tính là khác biệt có ý nghĩa thống kê với độ tin cậy 99.9% vì p–value của Kruskal–Wallis H test < 0.001. Kết luận là lượng SGPT trung bình của nhóm hôn mê gan do bệnh cấp tính là 71.421 lớn hơn một cách thực sự so với lượng SGPT trung bình (33.267) của nhóm hôn mê gan do bệnh mạn tính.

```
Lệnh:     MEANS SGPT TINHTHAN ↵
 TINHTHAN      Obs      Total       Mean     Variance     Std Dev
        1       11        462     42.000     1481.200      38.486
        2       44       3120     70.909     1929.712      43.928
        3       17        988     58.118     1473.485      38.386
 TINHTHAN    Minimum   25%ile    Median   75%ile  Maximum    Mode
        1     15.000   20.000    27.000   42.000  122.000  21.000
        2     17.000   35.500    72.000   96.500  210.000  72.000
        3     12.000   27.000    59.000   78.000  143.000  12.000
                                ANOVA
                (For normally distributed data only)
Variation      SS          df       MS        F statistic     p-value
Between     7992.543       2     3996.272       2.272        0.110767
Within     121365.401     69      1758.919
Total      129357.944     71
           Bartlett's test for homogeneity of variance
  Bartlett's chi square = 0.551  deg freedom = 2  p-value = 0.759208
          The variances are homogeneous with 95% confidence.
If samples are also normally distributed, ANOVA results can be used.
           Kruskal-Wallis One Way Analysis of Variance
Kruskal-Wallis H (equivalent to Chi square) =       5.655
                      Degrees of freedom =           2
                                 p value =     0.059161
```

Lượng SGPT

Trạng thái tinh thần	n	trung bình	phương sai	độ lệch chuẩn
Tỉnh táo	11	42.000	1481.200	38.486
Tiền hôn mê	44	70.909	1929.712	43.928
Hôn mê	17	58.118	1473.485	38.386

Trị số p của Bartlett's test: p–value = 0.759208.

Trị số p của ANOVA test: p–value = 0.110767.

Phương sai của lượng SGPT ở ba nhóm của trạng thái tinh thần: Tỉnh táo, Tiền hôn mê, Hôn mê là khác biệt không có ý nghĩa thống kê (p = 0.759208). Lượng SGPT trung bình ở ba nhóm cũng khác biệt không có ý nghĩa thống kê (p = 0.110767).

2. LỆNH REGRESS – TƯƠNG QUAN HỒI QUY TUYẾN TÍNH (Linear regression)

Cú pháp: REGRESS <biến phụ thuộc> <biến độc lập>

Lệnh REGRESS – trong trường hợp đơn giản nhất – được dùng để tính tương quan hồi quy tuyến tính giữa biến phụ thuộc Y và biến độc lập X theo dạng phương trình Y=aX+b. Máy sẽ đưa ra kết quả tính hệ số tương quan r và hệ số tương quan bình phương r^2 (Correlation coefficient), hệ số a (coefficient) và hệ số b (Y – Intercept) của phương trình.

Ngoài ra lệnh REGRESS còn được dùng để tính hồi quy đa biến, nhưng ta không giới thiệu ở đây.

Ví dụ: Để tính tương quan hồi quy tuyến tính giữa lượng SGOT (biến X) và lượng SGPT (biến Y), ta

dùng lệnh:

```
SELECT SGPT>0 AND SGOT>0 ↵
REGRESS SGPT SGOT ↵
```

Được kết quả sau:

```
coefficient:              r   = 0.79   r^2 = 0.62
95% confidence limits:          0.45 < r^2 < 0.74
Source          df    Sum of Squares   Mean Square   F-statistic
Regression      1       66335.9083     66335.9083     110.96
Residuals       69      41249.2466     597.8152
Total           70      107585.1549
B Coefficients
                              B  95% confidence          Partial
Variable  Mean  coefficient  Lower    Upper     Std Error    F-test
SGOT    47.5761 1.3727380  1.112760 1.632716    0.130316   110.9639
Y-Intercept       -3.9010092
```

Vậy SGOT và SGPT (n=70) có mối tương quan tuyến tính chặt chẽ vì hệ số tương quan r = 0.79 hay r^2=0.62 và khoảng tin cậy 95% của r^2 không chứa "0". Đồng thời khoảng tin cậy 95% của hệ số a (hệ số của biến SGOT) là [1.112760 – 1.632716] không chứa giá trị "0", cũng chứng tỏ phương trình chúng ta đưa ra là có ý nghĩa thống kê. Hệ số a=1.3727380, hệ số b=–3.9010092. Ta có thể biểu diễn mối liên hệ giữa SGPT và SGOT bằng phương trình SGPT = 1.37 x SGOT – 3.90. Từ phương trình này khi biết lượng SGOT của một bệnh nhi hôn mê gan ta có thể ước lượng gần đúng giá trị SGPT của bệnh nhi đó.

Trong trường hợp r và r^2 nhỏ hay là khoảng tin cậy 95% của r^2 chứa giá trị "0", lúc đó ta cũng thấy khoảng tin cậy 95% của hệ số a cũng chứa giá trị "0" như ở ví dụ dưới đây khi tính tương quan tuyến tính giữa BC (bạch cầu) và TC (tiểu cầu). Kết luận là không có mối tương quan tuyến tính giữa BC và TC (n=35), ta dùng lệnh:

```
SELECT BC>0 AND TC>0 ↵
REGRESS BC TC ↵
```

Ta được kết quả:

```
Correlation coefficient:      r   = 0.26   r^2 = 0.07
95% confidence limits:   -0.27 < r^2 < 0.39

Source          df   Sum of Squares   Mean Square   F-statistic
Regression       1        246.4063      246.4063          2.42
Residuals       34       3458.5501      101.7221
Total           35       3704.9564
B Coefficients
                    B       95% confidence              Partial
Variable  Mean   coefficient  Lower    Upper    Std Error  F-test
TC      151.9444 0.0297580  -0.009099 0.068615  0.019120  2.4223
Y-Intercept      12.3478816
```

CÂU HỎI LƯỢNG GIÁ

1. Để xét mối tương quan giữa một biến định lượng và một biến định tính, ta chọn lệnh nào trong các lệnh sau:

a) VARIABLES
b) TABLES
c) SELECT
d) MEANS

2. Để so sánh lượng hồng cầu (HC) giữa 2 giới (GIOI), hãy chọn câu lệnh đúng:

a) TABLES HC GIOI
b) TABLES GIOI HC
c) MEANS GIOI HC
d) MEANS HC GIOI

3. Cú pháp lệnh MEANS <biến1> <biến2>, biến 1 và biến 2 phải là:

a) Biến 1 và biến 2 là biến định lượng.
b) Biến 1 và biến 2 là biến định tính.
c) Biến 1 là biến định tính, biến 2 là biến định lượng.
d) Biến 1 là biến định lượng, biến 2 là biến định tính.

4. Để tìm được P so sánh phương sai và P so sánh trung bình, sử dụng lệnh nào trong các lệnh sau:

a) FREQ
b) TABLES
c) REGRESS
d) MEANS

BÀI TẬP VÀ THỰC HÀNH

Bài 4.1: Trong tập tin VIEMGAN.REC, hãy:

a) Tính trung bình, phương sai và độ lệch của lượng SGPT, SGOT, BLTP, BLTT.
b) So sánh SGPT của hai nhóm bệnh nhân có xuất huyết và không có xuất huyết.
c) So sánh SGOT của hai nhóm bệnh nhân hôn mê gan do bệnh cấp tính và hôn mê gan do bệnh mạn tính.
d) Tính tương quan giữa BLTP và BLTT.

Bài 4.2: Theo dõi dấu hiệu viêm khớp khi điều trị (ĐT) bệnh nhân viêm đa khớp thu được số liệu sau:

Trước ĐT	3	2	6	4	7	12	5	4	8	15	18	15
Sau 1 tháng ĐT	3	2	4	4	6	10	5	4	8	14	18	13
Sau 2 tháng ĐT	2	0	4	2	4	7	3	2	4	10	15	7

(tiếp)Trước ĐT	20	16	8	15	17	16	18	15	9	13
(tiếp) Sau 1 tháng ĐT	18	15	9	14	15	14	20	15	8	12
(tiếp) Sau 2 tháng ĐT	15	13	7	8	10	12	17	13	7	10

a) Tính các tham số $\bar{x} \pm s$ của 3 dãy số liệu: trước điều trị, sau 1 tháng ĐT, sau 2 tháng ĐT.

b) Tính các tham số $\bar{z} \pm s$ của chênh lệch trước ĐT và sau 1 tháng ĐT, của trước ĐT và sau 2 tháng ĐT, của sau 1 tháng ĐT và sau 2 tháng ĐT.

c) Hãy so sánh từng cặp của số khớp viêm trước ĐT và sau 1 tháng ĐT, của trước ĐT và sau 2 tháng ĐT, của sau 1 tháng ĐT và sau 2 tháng ĐT để đánh giá hiệu quả của phương pháp điều trị.

Bài 4.3: Điều trị sốt rét bằng 4 cách. Theo dõi thời gian hết KST sốt rét trong máu (giờ) của từng bệnh nhân thu được số liệu sau:

Cách 1	18	37	46	46	46	50.5	61.5	78	84.5	90
Cách 2	38	41	41.1	42	43.1	44.1	45.2	50	50	52
Cách 3	36	48	50	52	58	60	60	68	74	74
Cách 4	36	38	40	42	48	60	62	70	72	72

a) Tính các tham số $\bar{x} \pm s$ của từng cách điều trị.

b) Hãy so sánh trung bình của 2 trong 4 cách với nhau.

c) Hãy so sánh 4 giá trị trung bình của 4 cách điều trị.