**TÓM TẮT DỮ LIỆU CHO MỘT BIẾN ĐỊNH LƯỢNG**

**Phân phối tần số**

Ba bước để xác định các nhóm cho một phân phối tần số:

1. Xác định số lượng các nhóm riêng biệt.
2. Xác định độ rộng của nhóm.
3. Xác định các giới hạn của nhóm.



**Số nhóm**: Nhóm được hình thành bằng cách xác định một phạm vi để nhóm các dữ liệu.

Nên sử dụng từ 5 đến 20 nhóm.

Nếu có ít phần tử thì nên sử dụng năm hay sau nhóm để tổng hợp dữ liệu.

Mục đích là sử dụng đủ các nhóm để cho thấy sự thay đổi trong các dữ liệu.

**Độ rộng của các nhóm**:

Độ rộng của mỗi nhóm là như nhau.

Nhiều nhóm có nghĩa là độ rộng của nhóm nhỏ hơn và ngược lại.

Để xác định độ rồng gần đúng của nhóm, sử dụng công thức:



Độ rộng của nhóm có thể được làm tròn sao cho thuận tiện khi xây dựng phân phối tần số. Ví dụ, độ rộng nhóm xấp xỉ 9,28 có thể được làm tròn lên 10.

**Giới hạn nhóm**: Giới hạn nhóm được lựa chọn sao cho mỗi giá trị quan sát thuộc về một và chỉ một nhóm.

*Giới hạn dưới* xác định giá trị nhỏ nhất có thể được gán cho nhóm.

*Giới hạn trên* xác định giá trị lớn nhất có thể được gán cho nhóm.

Ví dụ trong bảng 2.4, ta chọn giới hạn dưới cho nhóm là 10 và giới hạn trên là 14 cho nhóm đầu tiên trong bảng 2.5.

Giới hạn dưới là 15 và giới hạn trên là 19 cho nhóm tiếp theo.

Có tổng cộng năm nhóm: 10-14, 15-19, 20-24, 25-29 và 30-34.

Chênh lệch giữa giới hạn dưới của hai nhóm liền kề là độ rộng nhóm.

Bằng cách sử dụng hai giới hạn dưới đầu tiên của nhóm là 10 và 15, ta có độ rộng nhóm là 15 – 10 = 5.

Với số lượng các nhóm, độ rộng nhóm và giới hạn xác định nhóm, một phân phối tần số có thể thu được bằng cách đếm số giá trị thuộc mỗi nhóm.



Sử dụng bảng phân phối tần số 2.5 ta có thể rút ra những điều sau:

1. Thời gian kiểm toán thường gặp nhất là trong nhóm 15 – 19 ngày. Tám trong số 20 thời gian kiểm toán thuộc nhóm này.
2. Chỉ có một lần kiểm toán yêu cầu 30 ngày trở lên.

**Trị số giữa nhóm**: là trị số giữa của nhóm ở chính giữa giá trị thấp nhất và giá trị cao nhất. Với dữ liệu kiểm toán, trị số giữa của 5 nhóm này là 12, 17, 22, 27 và 32.

**Phân phối tần suất và tần suất phần trăm**

Tần suất là tỷ lệ của các quan sát thuộc về một nhốm với n quan sát.



Tần suất phần trăm của một nhóm là tần suất nhân 100.

Bảng phân phối tuần suất và tần suất phần trăm các nhóm trong bảng 2.5.



**Đồ thị điểm (Dot plot)**

Trục ngang cho thấy phạm vi biến thiên của dữ liệu. Mỗi giá trị dữ liệu được đại diện bởi một điểm đặt trên trục. Đồ thị 2.4 là đồ thị điểm cho dữ liệu thời gian kiểm toán trong bảng 2.4.



Ba dấu chấm nằm phía trên vị trí số 18 của trục ngang chỉ ra thời gian kiểm toán 18 ngày xảy 3 lần.

**Biểu đồ phân phối (Histogram)**



Minh hoạ 2.5 là biểu đồ phân phối cho dữ liệu thời gian kiểm toán. Ta thấy nhóm nhóm có tần số lớn nhất nằm ở nhóm 15 – 19 ngày, tần số nhóm là 8.

Biểu đồ phân sphối khác với biểu đồ thanh là không có khoảng trống giữa các thanh.

Một trong những ứng dụng quan trọng của biểu đồ phân phối là để cung cấp thông tin về hình dạng của một phân phối. Mô tả 2.6 có 4 biểu đồ được xây dựng từ bản phân phối tần suất.



**Hình A** biểu diễn một tập hợp các dữ liệu tương đối lệch sang bên trái. Điểm thi là ví dụ điển hình cho dạng biểu đồ này, không có điểm số trên 100%, hầu hết các điểm số trên 70% và chỉ một vài điểm thực sự thấp.

**Hình B** cho thấy biểu đồ của một tập hợp các dữ liệu tương đối lệch về phía bên phải. Một ví dụ của loại đồ thị này như giá nhà đất, một vài ngôi nhà đắt tiền sẽ tạo ra độ lệch ở đuôi phải.

**Hình C** cho thấy một biểu đồ đối xứng. Trong biểu đồ đối xứng, đuối trái phải chiếu hình dạng đuôi phải. Dữ liệu điểm SAT, chiều cao và trọng lượng của người dân thường là biểu đồ đối xứng.

**Hình D** cho thấy biểu đồ rất lệch về phía bên phải. Biểu đồ này được xây dựng từ dữ liệu về số lượng khách hàng mua hơn một ngày tại một cửa hàng quần áo. Dữ liệu từ các ứng dụng trong kinh doanh và kinh tế tường dẫn đến đồ thị lệch về bên phải. Ví dụ, dữ liệu và giá nhà ở, tiền lương, số tiền mua.

**Phân phối tích luỹ (Cumulative Distribution)**

Phân phối tần số tích luỹ sử dụng số lượng các nhóm, chiều rộng các nhóm và giới hạn nhóm đã xây dựng cho phân phối tần số.

Phân phối tần số tích luỹ cho thấy số lượng các dữ liệu có giá trị nhỏ hơn hoặc bằng với giới hạn trên của nhóm tương ứng.

Hai cột đầu tiên của bảng 2.7 cung cấp Phân phối tần số tích luỹ cho dữ liệu thời gian kiểm toán.



Xem xét các nhóm có giá trị ít hơn hoặc bằng 24. Tần số tích luỹ cho nhóm này đơn giản là tổng tất cả các nhóm có giá trị nhỏ hơn hoặc bằng 24. Đối với bảng phân phối tần số trong bảng 2.5, tổng các tần số cho các nhóm 10-14, 15-19 và 20-24 có giá trị là 17 và nhỏ hơn hoặc bằng 24. Do đó tần số tích luỹ cho nhóm này là 17.

Ngoài ra, phân phối tần số tích luỹ trong bảng 2.7 cho thấy rằng 4 lần kiểm toán đã được hoàn thành trong 14 ngày hoặc ít hơn 19 việc kiểm toán đã được hoàn thành trong 29 hoặc ít hơn.

**Phân phối tần suất tích luỹ** cho thấy tỷ lệ các dữ liệu và **phân phối tần suất phần trăm** cho thấy tỷ lệ phần trăm của các dữ liệu có giá trị nhỏ hơn hoặc bằng giới hạn của nhóm.

Phân phối tần suất tích luỹ được tính bằng cách cộng các tần suất trong Phân phối tần suất hoặc bằng cách chia tần số tích luỹ cho tổng các giá trị.

**Đồ thị Ogive**

Đồ thị Ogive cho thấy giá trị dữ liệu trên trục ngang và một trong số tần số tích luỹ, hoặc tần suất tích luỹ hoặc tần suất phần trăm tích luỹ trên trục thẳng đứng.



Bài tập:

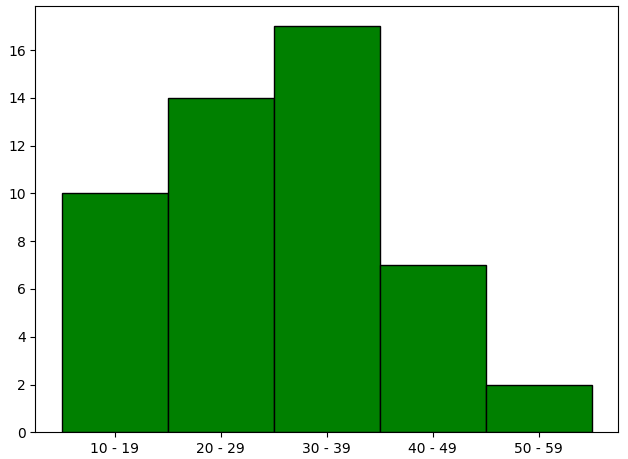
11)



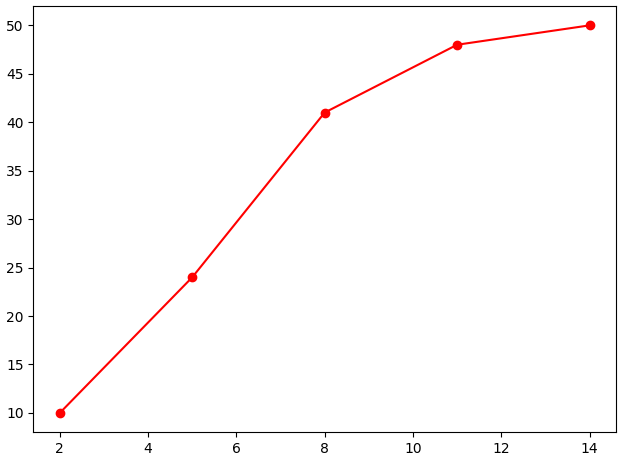
12)

13)

Đồ thị Histogram cho bảng phân phối ở bài 12:

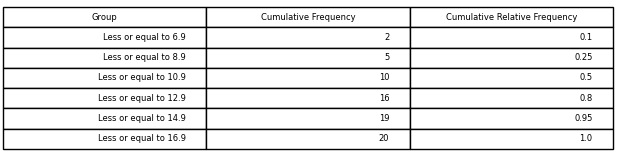


Đồ thị Ogive cho bảng phân phối bài 12:

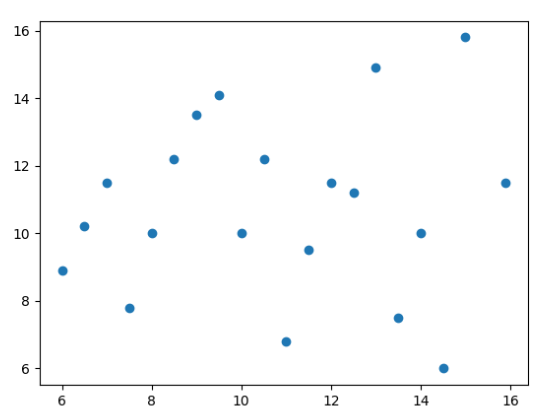


14)

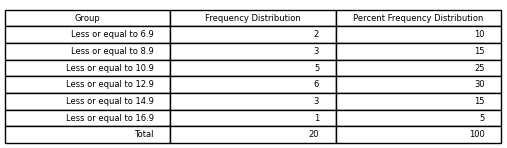
Bảng phân phối tần số tích lũy:



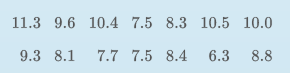
Đồ thị điểm của dữ liệu:

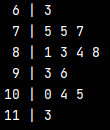


Bảng phân phối tần suất và phần trăm tần suất

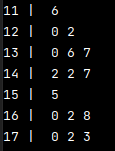


15) Đồ thị thân và nhánh cho dữ liệu sau:



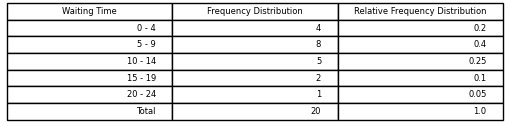


16) Biểu đồ thân và lá cho dữ liệu sau:

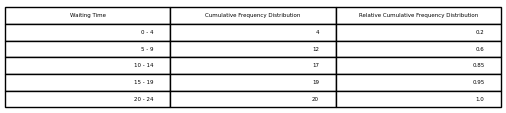


17)

a, b) Bảng phân phối tần số và tần suất của thời gian chờ của bệnh nhân:



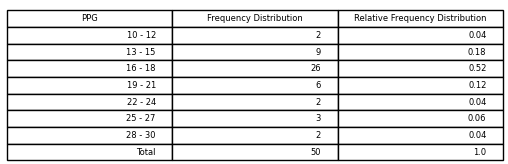
c, d) Bảng phần phối tần số tích lũy và tần suất tích lũy của thời gian chờ của bệnh nhân:



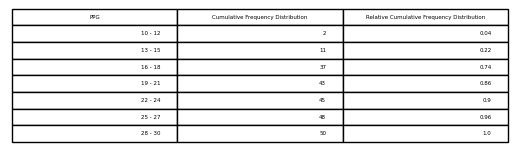
e) Tỷ lệ bệnh nhân chờ dịch vụ khẩn ít hơn 9 phút là 0.6.

18)

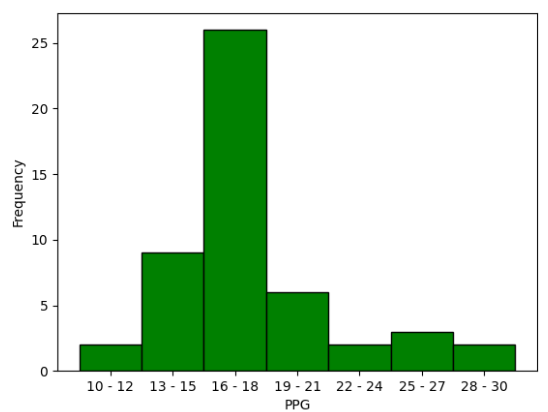
a, b)



c)

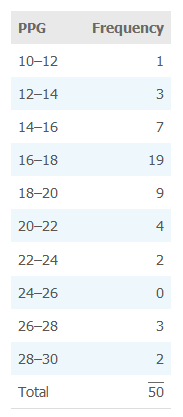


d) Biểu đồ Histogram cho điểm trung bình của mỗi game:



e) Biểu đồ bị lệch về phía bên phải.

d) Số điểm trung bình người chơi có số điểm ít nhất là 20 là 78%.

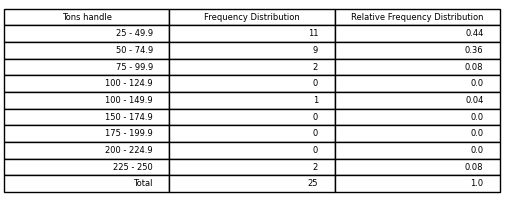


19)

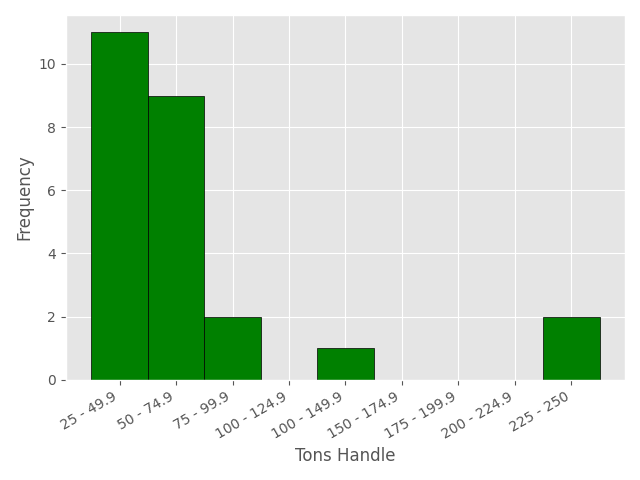
a) Số tấn phải xử lý lớn nhất là: 236.3

Số tấn phải xử lý nhỏ nhất là: 30.2.

b) Bảng phân phối tần số



c) Đồ thị histogram

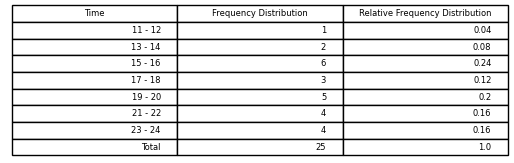


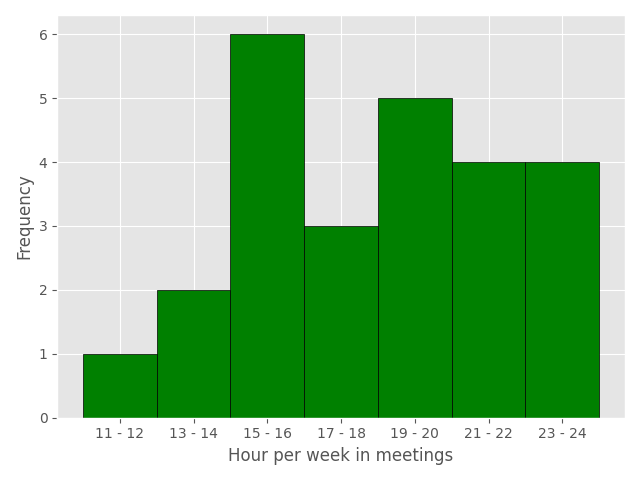
20)

a) Giá trị thời gian thấp nhất: 12 giờ

Giá trị thời gian lớn nhất: 23 giờ.

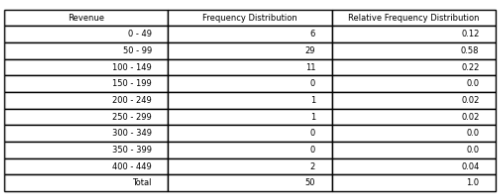
b) Bảng phân phối tần số và phân phối tần suất



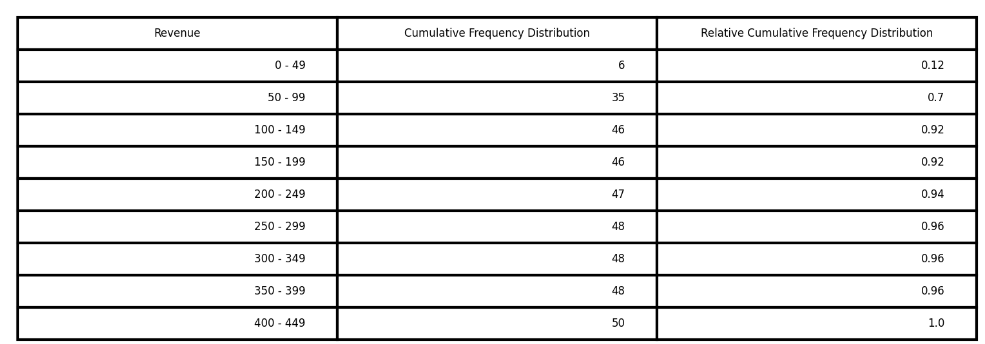
c) Đồ thị histogram

21)

a, b) Bảng phân phối tần số và tần suất:

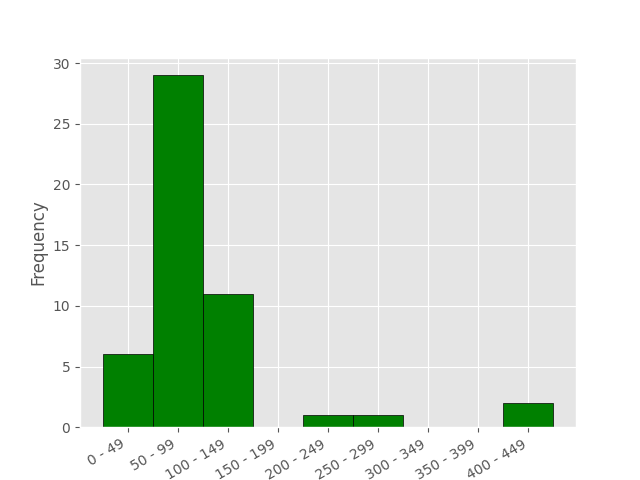


c, d) Bảng phân phối tần số và tần suất tích luỹ



e) Các tập đoàn có doanh thu từ 50 đến 99 tỉ $ chiếm phần lớn, có 11 tập đoàn có doanh thu từ 100 đến 149 tỉ $, có 6 tập đoàn có doanh thu thấp nhất từ 0 đến 49 tỉ $ và có 2 tập đoàn có doanh thu lớn nhất từ 400 đến 449 tỉ $.

f) Đồ thị histogram

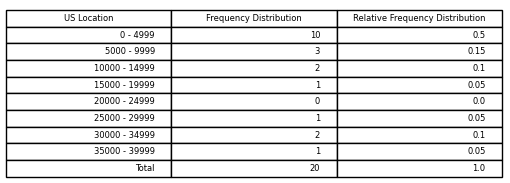


Đồ thị bị lệch nhiều về phía bên phải.

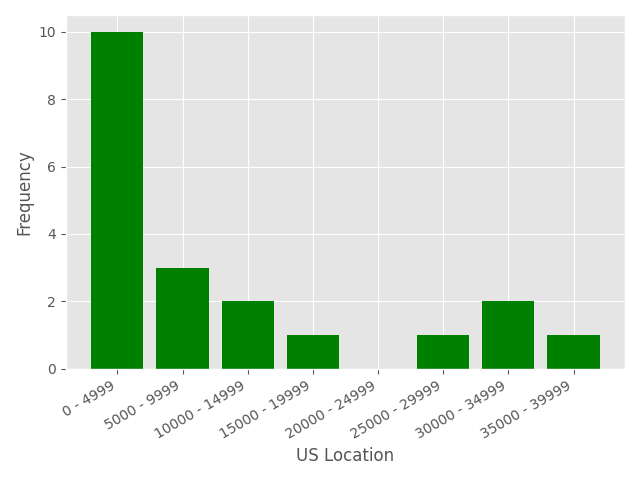
g) Tập đoàn lớn nhất ở Mĩ là Exxon Mobil và có doanh thu hàng năm là 443 tỉ $.

22)

a) Bảng phân phối tần số và tần suất



b) Biểu đồ histogram

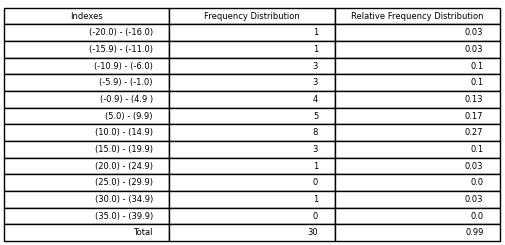


c) Đồ thị phân phối bị lệch phải, phần lớn các nhượng quyền thương mại trong danh sách đều ít hơn 20,000 địa điểm (50% + 15% + 15% = 80%). Mc Donald, Subway và 7-Eleven có số địa điểm cao nhất.

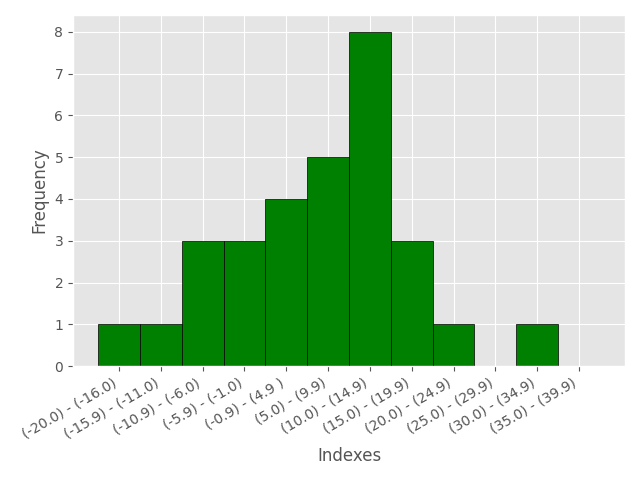
23)

a) Chỉ số Index có giá trị dương lớn nhất: Nikkei (31.4)

b) Bảng phân phối tần số



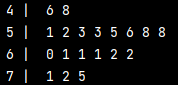
c)



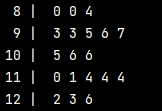
Đồ thị đối xứng nhưng hơi lệch phải.

24)

Đồ thị stem-leaf của mức lương trung bình khởi điểm:

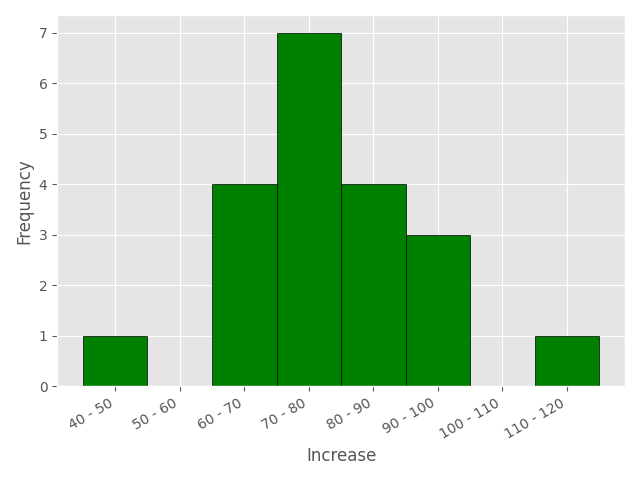


Đồ thị stem-leaf của mức lương trình giữa sự nghiệp



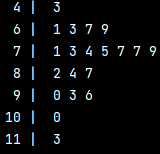
25)

a) Đồ thị histogram về mức tăng phần trăm của lương khởi điểm



b) Đồ thị có kiểu dáng gần như đối xứng nhưng lệch một ít về bên trái.

c) Đồ thị thân-nhánh về mức tăng phần trăm của lương khởi điểm

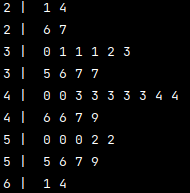


d) Ưu điểm của đồ thị stem-leaf so với histogram

http://www.netmba.com/statistics/plot/stem/

26)

a) Đồ thị thân-lá kéo dài:



b) Nhóm tuổi có số người chạy nhiều nhất: 40 – 44 có 9 người chạy.

c) Độ tuổi có người chạy thường xuyên: 43 có 5 người chạy.