TEST 2: Chapter 4,6,7

Chapter 3 là biến ngẫu nhiên rời rạc thì chapter này nói về loại biến còn lại: biến liên tục. Nó sẽ hơi khó hơn 1 chút so với biến rời rạc, nặng về toán hơn và cũng yêu cầu tư duy tốt hơn. Chúng cũng có các loại tham số và loại hàm như biến rời rạc: hàm phân bố(density function), hàm tích lũy (Cumulative), kỳ vọng(mean), phương sai (variance), độ lệch chuẩn (standard deviation) . Và sau đó ta sẽ được tiếp cận với loại khái niệm mới sẽ đi cùng xuyên suốt đến hết môn học: phân bố chuẩn (normal distributions).

Chapter 4: Continuous Random Variables and Probability Distribution

(Biến ngẫu nhiên liên tục và phân bố xác suất)

4.1 Continuous Random Variables (biến ngẫu nhiên liên tục)

Just a definition:
Biến ngẫu nhiên liên tục (Continuous Random Variables) là biến ngẫu nhiên có một khoảng các số thực cho các khả năng của nó.

VD: tốc độ, xe bạn có thể đi trong khoảng 0-50km/h, vậy bạn có thể đạt bất kì giá trị nào, 40.00001, 25, 36.9999, ...

4.2 Probability Distributions and Probability Density Functions (phân bố xác suất và hàm mật độ xác suất)

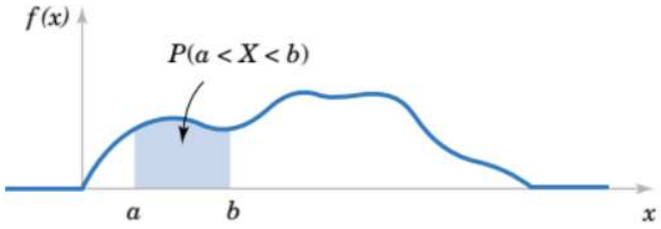
f(x) là hàm phân bố xác suất của biến ngẫu nhiên X nếu :

$$(1)\ f(x) \geq 0$$

$$(2)\ \int_{-\infty}^{\infty} f(x)dx = 1$$

$$(3)\ P(a < X < b) = \int_{a}^{b} f(x)dx \text{ for any } a \text{ and } b.$$

Đây là hàm phân bố, nó liên tục chứ không giống với biến rời rạc. Đối với biến rời rạc, mỗi giá trị của X nó sẽ có giá trị cụ thể của X kèm theo nó. Nhưng với biến liên tục, nó sẽ có vô hạn giá trị nên xác suất để xảy ra giá trị cụ thể sẽ = 0. Vậy nên ta thường tính xác suất theo 1 khoảng, nó sẽ bằng tích phân (hay diện tích) trong khoảng đó.



Như hình trên ta thấy xác suất của X để nó nằm trong khoảng a và b là tích phân từ a đến b, hay diện tích trong khoảng đó.

$$\text{i)}\ P(X = a) = 0,$$
$$\text{ii)}\ P(a \le X \le b) = P(a \le X < b) = P(a < X \le b) = P(a < X < b)$$

Chú ý 1 điều rằng dấu "<" và "≤" tương đương nhau trong biến liên tục, bởi xác suất để xảy ra giá trị cụ thể luôn = 0. Bạn thử tưởng tượng khi đi xe, bạn không thể nào đi đúng 40Km/h được, bởi có thể bạn đang đi 39.99999 km/h hay 40.0001km/h. Chẳng có gì là tuyệt đối cả.

4.3 Cumulative Distribution Functions (Hàm phân bố tích lũy)

Hàm phân bố tích lũy:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt,\ \forall x \in \mathbb{R}$$

Cũng giống như bên phần rời rạc thì phần liên tục cũng có concept tương tự, là tại giá trị x thì F(x) sẽ có giá trị là xác suất để ≤ x. ta thấy nó sẽ là tích phân từ $-\infty$, nhưng nếu đề cho khoảng giá trị ban đầu là a≤X≤b thì ta có thể thay $-\infty$ bằng a để tính F(x).

$$P(a < X < b) = F(b) - F(a)$$

Công thức trên khá dễ để hiểu và cũng hay dùng, giá trị của F(b) bằng xác suất để X≤b, F(a) là xác suất để X≤a, vậy thì xác suất để a≤X≤b sẽ là F(b)-F(a).

4.4 Mean and Variance of a Continuous Random Variable (Kỳ vọng và phương sai của biến liên tục)

f(x) là density function (hàm phân bố), ta có:

mean (kỳ vọng) :

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

variance (phương sai) :

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

standard deviation (độ lệch chuẩn) :

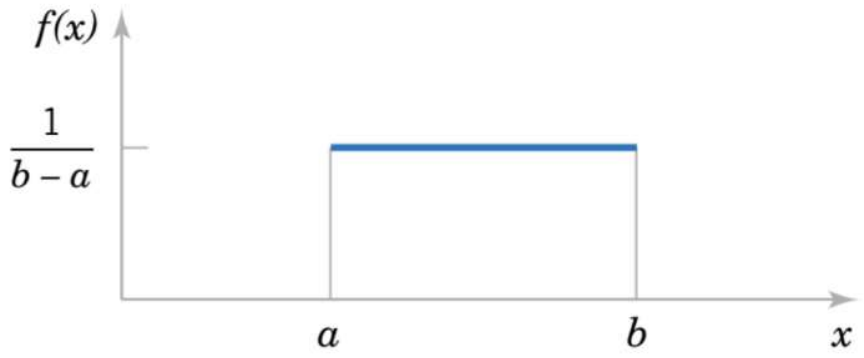
$$\sigma = \sqrt{\sigma^2}$$

4.5 Continuous Uniform Distribution (phân bố đồng đều liên tục)

Cũng tương tự như biến rời rạc, biến liên tục cũng có phần đồng đều, khá dễ và chúng có dạng:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases}$$

Đồ thị:



mean (kỳ vọng) : và variance (phương sai) :

$$\mu = E(X) = \frac{a+b}{2},\ \sigma^2 = V(X) = \frac{(b-a)^2}{12}$$

Cumulative function (Hàm tích lũy) của biến liên tục phân bố đồng đều:

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$$

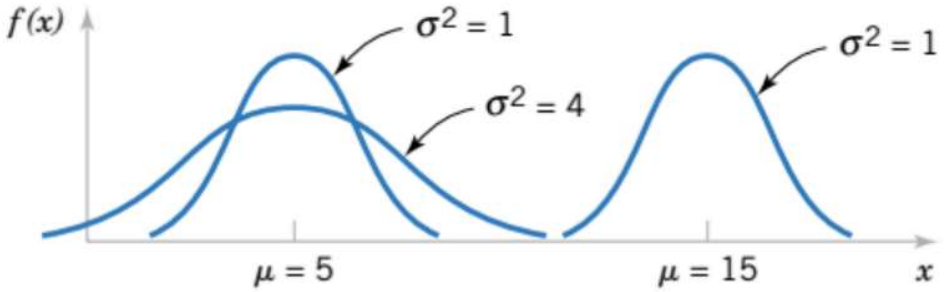
4.6 Normal Distribution (Phân bố chuẩn)

Phần này quan trọng, nó sẽ liên quan đến nhiều thứ về sau. Nó sẽ sử dụng cho các bài toán mang tính chất có lượng dữ liệu rất lớn. Sau này ta sẽ thấy họ sử dụng cái này để ước lượng các thông số của tổng thể (Population) thông qua một phần dữ liệu (Sample).
VD: nghiên cứu lương 1000 người dân để đánh giá lương trung bình của toàn dân số.

Biến ngẫu nhiên X sẽ được coi là có phân bố chuẩn (normal distribution) với tham số μ và σ², ký hiệu X ~ N(μ, σ²), nếu như hàm phân bố xác suất của nó là:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Tin vui là cái hàm trên bạn không phải nhớ, chỉ là giới thiệu qua thôi. Để dễ hình dung hơn về mean μ và variance σ² , ta quan sát biểu đồ sau:

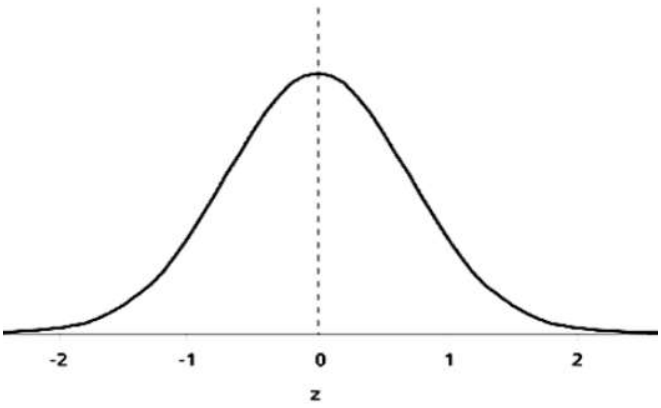


Ta thấy rằng μ chính là kỳ vọng, giá trị trung bình của Population(tổng thể), các giá trị sẽ tập trung nhiều về đó, còn σ² biểu diễn cho sai số nên dễ dàng thấy rằng phương sai càng lớn thì đồ thì sẽ thoải hơn, không tập trung nhiều gần μ bằng khi σ² nhỏ. Và cái này sẽ giúp bạn hiểu bản chất hơn thôi, chứ thi thì không có phần đồ thị này.

mean (kỳ vọng) : và variance (phương sai) :

$$\text{If } X \sim N(\mu, \sigma^2), \text{ then } E(X) = \mu, V(X) = \sigma^2.$$

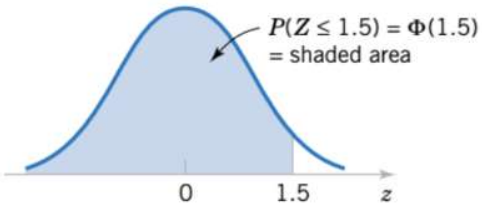
Standard Normal Distribution :
Từ đây, chúng ta sẽ không dùng biểu đồ như bên trên nữa, mà mọi bài toán liên quan đến phân bố chuẩn sẽ được quy về Z. với các tham số μ = 0, σ = 1, nó kiểu như đồ thị phân bố của các độ lệch chia cho standard deviation (độ lệch chuẩn). Đồ thị nó sẽ như này:



Còn hàm Cumulative (tích lũy) của nó sẽ như sau:

$$\Phi(z) = P(Z \leq z).$$

Cái này dùng nhiều, nó là diện tích từ z trở về trước, là xác suất để Z ≤ z, VD:



Khi ta có giá trị μ , σ của X, thì ta sẽ chuyển sang Z bằng cách:

$$\text{If } X \sim N(\mu, \sigma^2) \text{ then } Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Ta sẽ tính xác suất khi X nhỏ hơn a bằng cách quy đổi sang giá trị của z, và đồ thị của Z luôn cố định cho mọi bài toán:

$$P(X < a) = P\left(Z < \frac{a - \mu}{\sigma}\right)$$

$$P(a < X < b) = P\left(Z < \frac{b - \mu}{\sigma}\right) - P\left(Z < \frac{a - \mu}{\sigma}\right)$$

4.7 Normal Approximation to the Binomial and Poisson Distributions

(Phân bố xấp xỉ chuẩn cho Binomial and Poisson)

Binomial and Poisson là 2 loại phân bố ta đã học ở chapter 3. Nhưng mà khi đó ta chỉ làm với những thông số nhỏ, bạn thử tưởng tượng với binomial, khi ta có rất nhiều phép thử thì ta không thể tính bằng các con số có số mũ vài chục vài trăm được, vì vậy họ sẽ sử dụng phân bố xấp xỉ chuẩn cho những trường hợp này.

Binomial Distribution:
Như đã học, Binomial Distribution sẽ có 2 thông số là n và p, lần lượt biểu thị cho số phép thử và xác suất ra "success" ở mỗi lần thử. Ở dạng đó ta có mean $\mu = np$ và variance $\sigma^2 = np(1 - p)$.

Poisson Distribution: Poisson cũng vậy, đôi khi ra sẽ gặp phải trường hợp có thông số rất lớn ta cũng sẽ phải sử dụng phân bố xấp xỉ chuẩn. Các thông số của Poisson sẽ là $\mu = \lambda$ và $\sigma^2 = \lambda$.

Các công thức xác suất cho 2 trường hợp này:

$$P(X \leq x) = P(X \leq x + 0.5) \approx P\left(Z \leq \frac{x + 0.5 - \mu}{\sigma}\right)$$

$$P(X \geq x) = P(X \geq x - 0.5) \approx P\left(Z \geq \frac{x - 0.5 - \mu}{\sigma}\right)$$

Nó sẽ được áp dụng hiệu quả nếu μ lớn hơn 5.

4.8 Exponential Distribution (Phân bố mũ)

Với Poisson, ta có λ là số lượng trung bình/ 1 khoảng thời gian hoặc không gian. Thì với phân bố mũ, nó sẽ tính về khoảng thời gian hay không gian giữa 2 lần suất hiện liên tiếp.
VD: Với Poisson, ta có λ = 5 cuộc gọi/ ngày, thì với phân bố mũ, ta sẽ quan tâm tới khoảng cách trung bình giữa 2 lần suất hiện liên tiếp, tức trung bình mean = 1/λ = 4.8 giờ.

Hàm phân bố xác suất (Probability density function) :

$$f(x) = \lambda e^{-\lambda x}, \ x \geq 0$$

Cái hàm này là f nhỏ, phân bố chứ không phải là hàm tính xác suất nhé. Nên nhớ là phần biến liên tục thì ta sẽ tính xác suất bằng cách tích phân trong 1 khoảng nếu muốn tính xác suất để X nó rơi vào khoảng đó.

mean (kỳ vọng) : và variance (phương sai) cho exponential distribution:

$$\mu = E(X) = \frac{1}{\lambda}, \ \sigma^2 = V(X) = \frac{1}{\lambda^2}$$

Chapter này khá đơn lẻ bởi nó ít liên kết nhất với các chapter khác. Bạn có một tập dữ liệu, cái tập đấy gọi là sample.

Chapter 6: Descriptive Statistics (thống kê mô tả)

6.1 Numerical Summaries of Data

Ta có n observations(quan sát) trong sample là x1, x2, x3, ... xn.

Sample mean:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Kỳ vọng của sample đơn giản là trung bình của các giá trị.

Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

Phương sai là tổng bình phương sai số chia cho n-1. Nhưng ấn được máy tính nên không cần nhớ cái cồng kềnh này nhé.

Sample standard deviation ký hiệu là s.

Sample range : $r = \max(x_i) - \min(x_i).$

6.2 Stem-and-Leaf Diagrams

Khi có nhiều dữ liệu, họ sẽ chia thành cài bảng này cho dễ nhìn hơn. VD:

stem	leaf
10	15
11	058
12	01346
14	67
15	1

Giải: Dữ liệu mà ta có sẽ là : 101, 105, 110, 115, 118, ... Nói chung là cột 'leaf' thì là 1 chữ số, sau đó nối với bên 'stem'.

- sample median (khác với sample mean) là giá trị nằm giữa của các dữ liệu này. Nếu dữ liệu có n số thì sample median sẽ là số thứ (n+1)/2. nếu (n+1)/2 mà ra dạng x,5 thì sample median sẽ = trung bình số thứ x và số thứ x+1.
VD1: cho sample: 1, 3, 5, 6, 8, 9, 10.
Giải: Số số hạng là n=7, vậy thì (n+1)/2 = 4 suy ra sample median = số thứ 4 là 6.
VD2: cho sample: 2, 3, 4, 6, 7, 8.
Giải: Số số hạng là n=6, vậy thì (n+1)/2 = 3.5 suy ra sample median = trung bình số thứ 3 và thứ 4 là (4+6)/2 = 5.

- sample mode là giá trị xuất hiện nhiều nhất, nếu tất cả các số đều có số lượng như nhau thì không có sample mode

- Quartiles:
Ta chia data thành 4 phần bằng nhau thì đó gọi là Quartiles, có 3 điểm là q1, q2 (chính là median), q3. Xấp xỉ 25% số lượng observations ở dưới q1, 50% dưới q2 và 75% dưới q3. Cách tính q1: số thứ (1+n)/4 , nếu ra .5 thì lấy 2 số gần đó nhất chia trung bình
Cách tính q2: là median.
Cách tính q3: số thứ (1+n) x 3/4 , nếu ra .5 thì lấy 2 số gần đó nhất chia trung bình

- interquartile range: IQR = q3-q1.

6.3 Frequency Distributions and Histograms (Kiểu theo tần số, số lần xuất hiện hay tỉ lệ ấy)

Relative frequency distribution Ví dụ:

Number of Pets	Frequency	Relative Frequency	
1	150	37.5%	← 150/400 = 37.5%
2	90	22.5%	← 90/400 = 22.5%
3	110	27.5%	← 110/400 = 27.5%
4	30	7.5%	← 30/400 = 7.5%
5	20	5.0%	← 20/400 = 5.0%

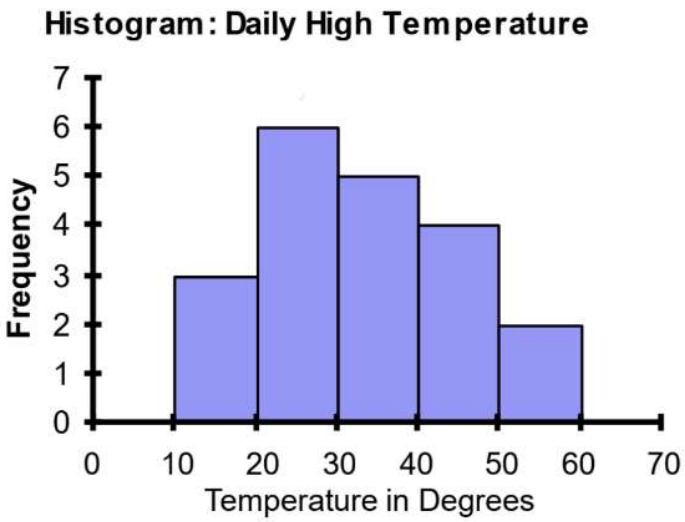
Cumulative frequency distribution ví dụ:

Type	Freq	cumulative frequency
Up to 1000	22	22
1001-2000	45	67
2001-3000	57	124
3001-4000	97	221
4001-5000	152	373
5001-6000	241	614

Nhớ lại: Cumulative là tích lũy, nó sẽ bằng tổng từ cái nhỏ nhất đến nó.

Histogram ví dụ:

Class	Frequency
10 – under 20	3
20 – under 30	6
30 – under 40	5
40 – under 50	4
50 – under 60	2



6.4 Box plot

Như phần 6.2, bạn đã biết được p1, p2, p3 là gì rồi. Và giờ ta dùng chúng để kiểu sàng lọc những phần tử quá khác biệt vậy. Box plot có 2 đầu mút. Bây giờ tính q3 + 1.5IQR, rồi lấy số lớn nhất trong data mà nhỏ hơn giá trị đấy làm đầu mút trên, sau đó tính q1-1.5IQR, lấy số nhỏ nhất trong data mà lớn hơn số đấy làm đầu mút dưới, vậy là đã có box plot, nếu trong data có số nào không nằm trong đó thì gọi là outlier.

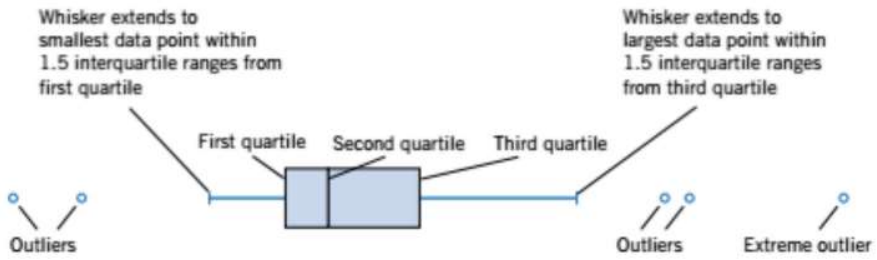


Figure: Description of a box plot

Chapter 7: Sampling Distributions and Point Estimation of Parameters

7.1 Introduction

Ta chọn ra X1, X2, X3, ... , Xn và gọi tập này là random sample với kích thước n.
VD: chọn 100 cái điện thoại từ nhà máy để test.

Sample mean và sample variance gọi là Statistic(bởi nó là của sample):

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1}$$

Nên nhớ đây là của Sample nhé, và chúng gọi là Statistic được kí hiệu như trên. Chứ của Population thì gọi là Parameter được ký hiệu là μ và σ².

Point Estimation

- Các Sample mean và sample variance là Statistic, ta gọi chúng lần lượt là point estimator của mean μ và variance σ² của Population.

- Còn khi tính ra số cụ thể, ta gọi nó là point estimate

7.2 Sampling Distributions and the Central Limit Theorem

Central Limit Theorem

Một Population có các tham số là μ và σ², và ta có sample mean $\overline{x}$. Bạn nhớ lại phần 4.6 khi ta học standard normal distribution, thì cách tính Z sẽ là

$$\text{If } X \sim N(\mu, \sigma^2) \text{ then } Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

.

Còn bây giờ ta chỉ tính cho sample, với n là số số hạng trong sample ta có:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

- Nếu như Population có mean μ và variance σ² thì khi ta lấy ra 1 sample, nó sẽ có các thông số :

$$\mu_{\overline{X}} = \mu, \ \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- Nếu Population là phân bố chuẩn, thì sample cũng là phân bố chuẩn.
- Nếu Population không phải phân bố chuẩn, thì sample sẽ là phân bố xấp xỉ chuẩn nếu như kích thước(lượng dữ liệu) ≥ 30.

Sampling Distribution of a Difference in Sample Means

Cái này là khi ta so sánh 2 cái sample với nhau. Ví dụ so sánh độ chênh lệch tuổi thọ chó với mèo chẳng hạn, thì ta sẽ có:

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$$

$$\sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$