

Statistical Basis for Bioinformatics

國立陽明大學 生物資訊研究中心

彭成煌(C.H. Perng)

chperng@ym.edu.tw

2002/09/30

Outline

Introduction

Data description

Statistical distribution

Statistical inference

Cluster v.s. Classification

Summary

Introduction

Example --- Microarray data

Question

Basic statistics

An example of microarray data

	A	B	ALL					AML				
1	Gene Description	Gene Accession Number	1	2	...	26	27	28	29	...	37	38
2	AFFX-BioB-5_at	AFFX-BioB-5_at	-214	-139	...	-112	-273	-4	15	...	-25	-72
3	AFFX-BioB-M_a	AFFX-BioB-M_at	-153	-73	...	-233	-327	-116	-114	...	-20	-139
4	AFFX-BioB-3_at	AFFX-BioB-3_at	-58	-1	...	-78	-76	-125	2	...	124	-1
5	AFFX-BioC-5_at	AFFX-BioC-5_at	88	283	...	54	81	241	193	...	325	392
6	AFFX-BioC-3_at	AFFX-BioC-3_at	-295	-264	...	-244	-439	-191	-51	...	-396	-324
7	AFFX-BioDn-5_a	AFFX-BioDn-5_at	-558	-400	...	-275	-616	-411	-155	...	-464	-510
8	AFFX-BioDn-3_a	AFFX-BioDn-3_at	199	-330	...	-479	419	-31	29	...	-221	-350
9	AFFX-CreX-5_at	AFFX-CreX-5_at	-176	-168	...	-108	-251	-240	-105	...	-390	-202
10	AFFX-CreX-3_at	AFFX-CreX-3_at	252	101	...	136	165	150	42	...	-1	249
11	AFFX-BioB-5_st	AFFX-BioB-5_st	206	74	...	-86	350	24	524	...	358	561
12	AFFX-BioB-M_st	AFFX-BioB-M_st	-41	19	...	-190	-40	19	-70	...	-197	275
13	AFFX-BioB-3_st	AFFX-BioB-3_st	-831	-743	...	-308	-1250	-669	-344	...	-852	-785
14	AFFX-BioC-5_st	AFFX-BioC-5_st	-653	-239	...	-497	-863	-664	-285	...	-689	-326
15	AFFX-BioC-3_st	AFFX-BioC-3_st	-462	-83	...	-106	-125	-311	-53	...	-175	47
16	AFFX-BioDn-5_s	AFFX-BioDn-5_st	75	182	...	80	270	297	38	...	-7	-96
17	AFFX-BioDn-3_s	AFFX-BioDn-3_st	381	164	...	246	312	353	-142	...	-156	16
18	AFFX-CreX-5_st	AFFX-CreX-5_st	-118	-141	...	-7	-179	-96	-125	...	-47	-140
19	AFFX-CreX-3_st	AFFX-CreX-3_st	-565	-423	...	-194	-874	-291	-118	...	-397	-367

T. R. Golub et al. (1999) *Science*, **286**, 531-537.

1. The datasets contain measurements corresponding to acute lymphoblastic leukemia(ALL) and acute myeloid leukemia(AML) samples from Bone Marrow and Peripheral Blood.
2. There are two datasets containing the training(38 samples, 27 ALL & 11 AML) and test(34 samples , 20 ALL & 14 AML).

Basic Statistics(1)

Gene	Mean	Median	Std	C.V.
1	-136.0	-112.0	109.6	-0.806
2	-158.0	-147.0	77.6	-0.491
30	-43.1	-49.0	39.9	-0.924
637	-165.5	4.0	583.6	-3.526
2940	77.4	-1.0	225.9	2.918
3000	148.1	161.0	82.4	0.556
5449	256.4	-6.0	744.8	2.905
6879	239.3	5.0	553.3	2.312
6923	127.5	3.0	519.0	4.070
7017	43.3	1.0	115.7	2.669

Data Description

Statistical measures

Graphical methods

Statistical tables

Other

Statistical Measures

Observations: X_1, X_2, \dots, X_n

Order Statistics: $X_{(1)}, X_{(2)}, \dots, X_{(n)}$

Location measures,
or Measures of central tendency

Dispersion measures,
or Measures of variability

Location

1. 平均數(Mean) $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ (統計量)

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i \quad (\text{參數})$$

2. 中位數(Median) $Md = \begin{cases} X_{(\frac{n+1}{2})} & , n \in \text{odd} \\ \frac{1}{2} [X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}] & , n \in \text{even} \end{cases}$

3. 眾數(Mode) Mo

4. 百分位數(Percentile), 第 k 百分位數(k -th Percentile)

$$P_k = \begin{cases} X_{([i]+1)} & , i \notin Z \\ \frac{1}{2} [X_{(i)} + X_{(i+1)}] & , i \in Z \end{cases} \quad \text{where } i = \frac{k}{100} n$$

Location(cont.)

1⁰. Sample size $n=50$  $P_{25} = X_{(12+1)} = X_{(13)}$
 $P_{50} = \frac{1}{2}[X_{(25)} + X_{(26)}]$

2⁰. $P_{50} = \text{Md}$

3⁰. 四分位數(Quartile) $Q_1 = P_{25}$, $Q_2 = P_{50} = \text{Md}$, $Q_3 = P_{75}$

4⁰. 十分位數(Deciles) $D_1 = P_{10}$, $D_2 = P_{20}$, ... , $D_9 = P_{90}$

Dispersion

1. 全距(Range)

$$R = X_{(n)} - X_{(1)}$$

2. 四分位距(Interquartile-range)

$$IQR = Q_3 - Q_1$$

3. 四分位差(Quartile deviation)

$$Q.D. = IQR/2 (= Q_2 - Q_1 = Q_3 - Q_2, \text{對稱資料})$$

4. 平均絕對偏差(Mean Absolute Deviation)

$$MAD = \frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}|$$

5. 變異數(Variance)

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$$

Dispersion(cont.)

6.標準差(Standard Deviation)

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} = \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right]} \quad (\text{統計量})$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N X_i^2 - \mu^2} \quad (\text{參數})$$

7.變異係數(Coefficient of Variation)

$$C.V. = \frac{S}{\bar{X}} 100\% \quad (\text{統計量})$$

$$C.V. = \frac{\sigma}{\mu} 100\% \quad (\text{參數})$$

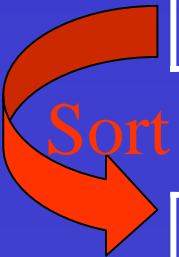
Basic Statistics

Gene	Mean	Median	Std	C.V.
1	-136.0	-112.0	109.6	-0.806
2	-158.0	-147.0	77.6	-0.491
30	-43.1	-49.0	39.9	-0.924
637	-165.5	4.0	583.6	-3.526
2940	77.4	-1.0	225.9	2.918
3000	148.1	161.0	82.4	0.556
5449	256.4	-6.0	744.8	2.905
6879	239.3	5.0	553.3	2.312
6923	127.5	3.0	519.0	4.070
7017	43.3	1.0	115.7	2.669

Example: Gene 3000(ALL)

Observations X_i

<i>Sample</i>	<i>01</i>	<i>02</i>	<i>03</i>	<i>04</i>	<i>05</i>	<i>06</i>	<i>07</i>	<i>08</i>	<i>09</i>
Intensity	310	204	234	271	233	-16	124	94	176
<i>Sample</i>	<i>10</i>	<i>11</i>	<i>12</i>	<i>13</i>	<i>14</i>	<i>15</i>	<i>16</i>	<i>17</i>	<i>18</i>
Intensity	148	36	87	211	222	189	113	217	153
<i>Sample</i>	<i>19</i>	<i>20</i>	<i>21</i>	<i>22</i>	<i>23</i>	<i>24</i>	<i>25</i>	<i>26</i>	<i>27</i>
Intensity	92	169	24	162	85	221	53	161	25



Order Statistics $X_{(i)}$

<i>Order</i>	<i>01</i>	<i>02</i>	<i>03</i>	<i>04</i>	<i>05</i>	<i>06</i>	<i>07</i>	<i>08</i>	<i>09</i>
Intensity	-16	24	25	36	53	85	87	92	94
<i>Order</i>	<i>10</i>	<i>11</i>	<i>12</i>	<i>13</i>	<i>14</i>	<i>15</i>	<i>16</i>	<i>17</i>	<i>18</i>
Intensity	113	124	148	153	161	162	169	176	189
<i>Order</i>	<i>19</i>	<i>20</i>	<i>21</i>	<i>22</i>	<i>23</i>	<i>24</i>	<i>25</i>	<i>26</i>	<i>27</i>
Intensity	204	211	217	221	222	233	234	271	310

Example: Gene 3000(ALL,cont.)

Order Statistics $X_{(i)}$

Order	01	02	03	04	05	06	07	08	09
Intensity	-16	24	25	36	53	85	87	92	94
Order	10	11	12	13	14	15	16	17	18
Intensity	113	124	148	153	161	162	169	176	189
Order	19	20	21	22	23	24	25	26	27
Intensity	204	211	217	221	222	233	234	271	310

Location	Min.	Q_1	Med.	Me.	Q_3	Max.
Value	-16	87(89.5)	161	148.5	217(214)	310
Index	1	6.75	13.5	*	20.25	27

Dispersion	R	IQR	Q.D.	Std	C.V.
Value	326	130(124.5)	65(62.25)	82.4	0.556

Graphical Methods

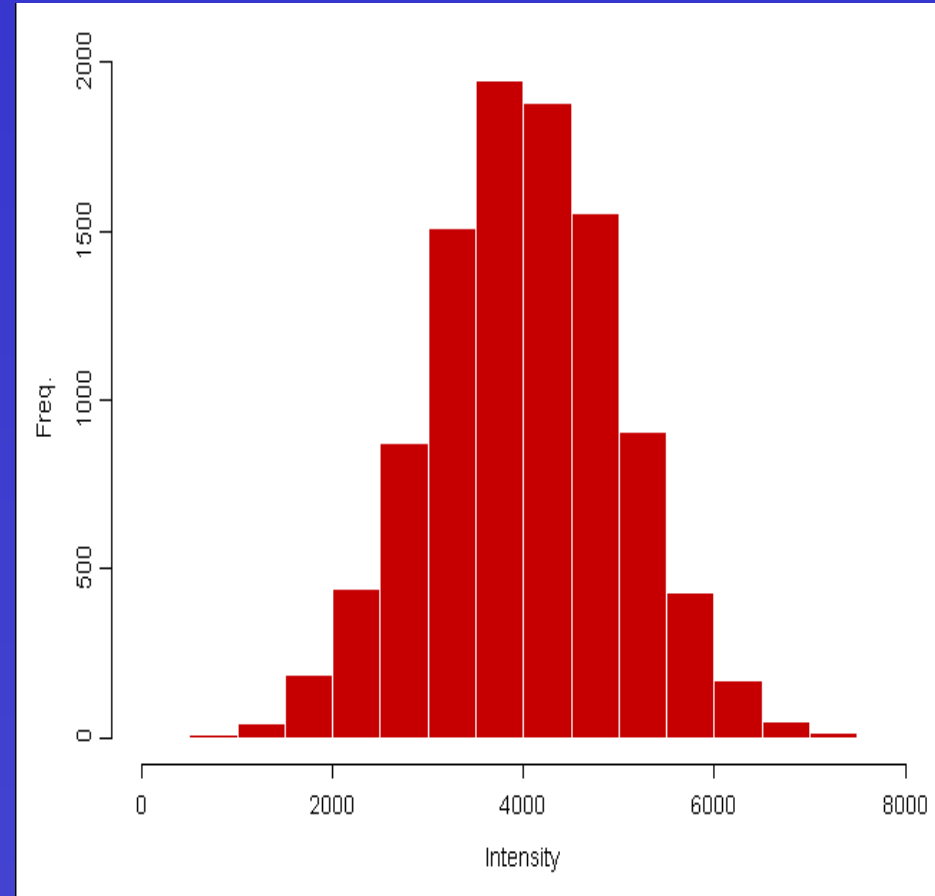
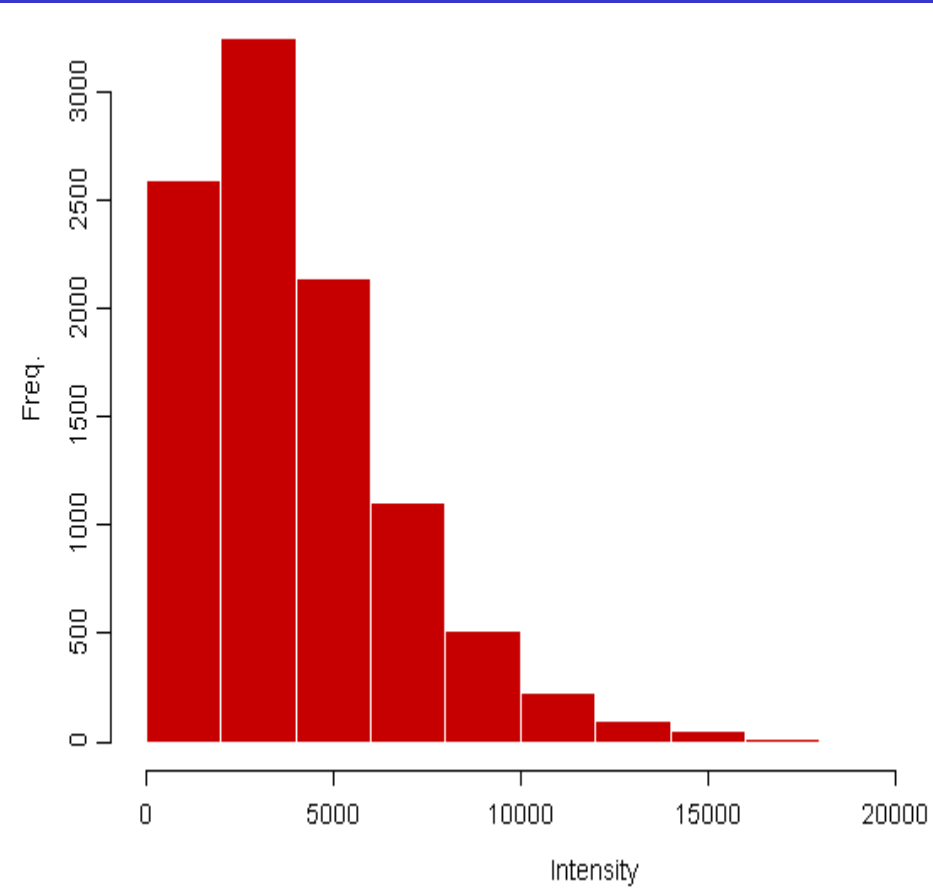
Histogram

Box-plot

Scatter plot

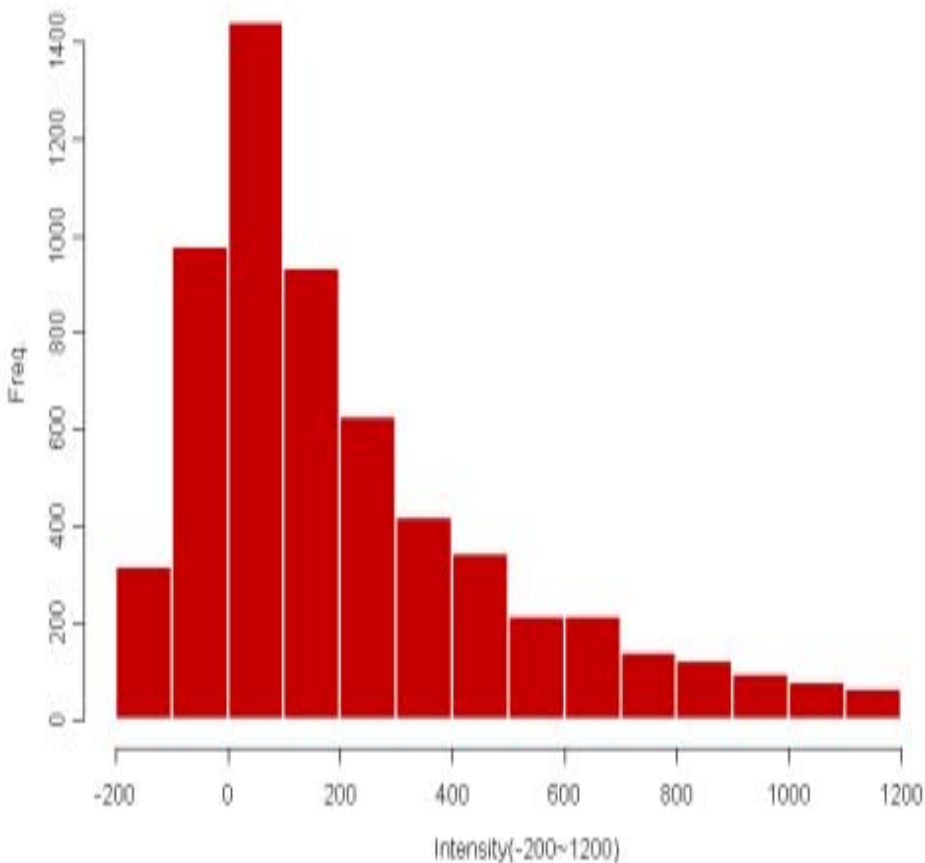
Other

Histogram

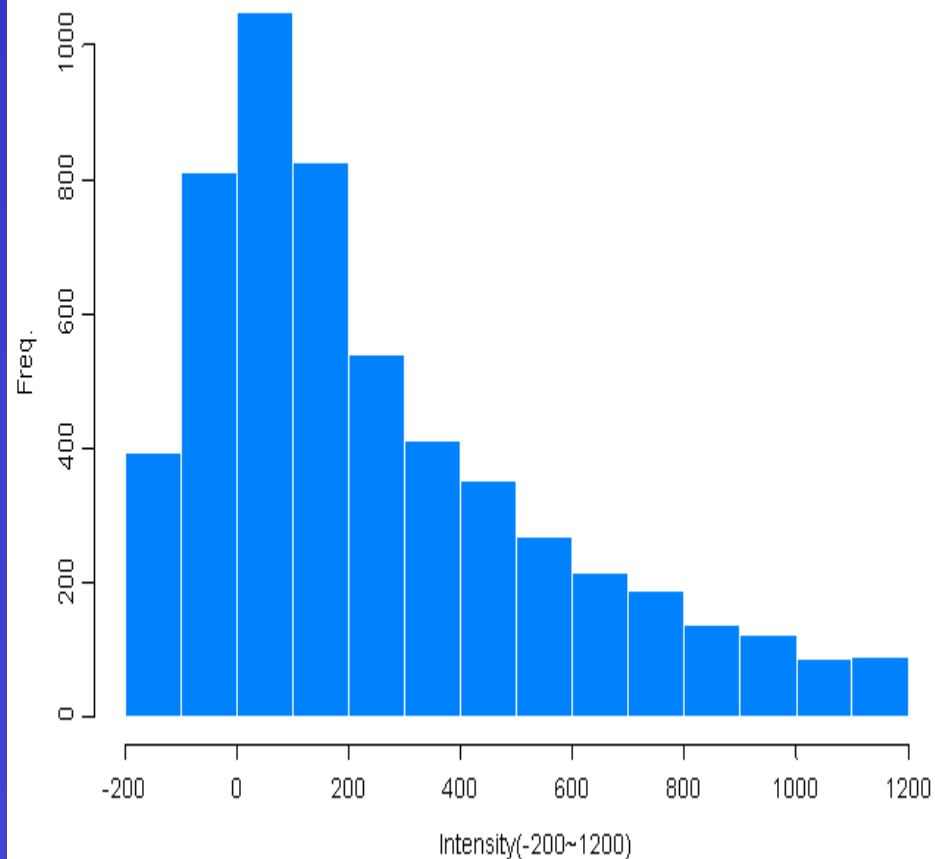


Example: ALL v.s. AML (Sample 5)

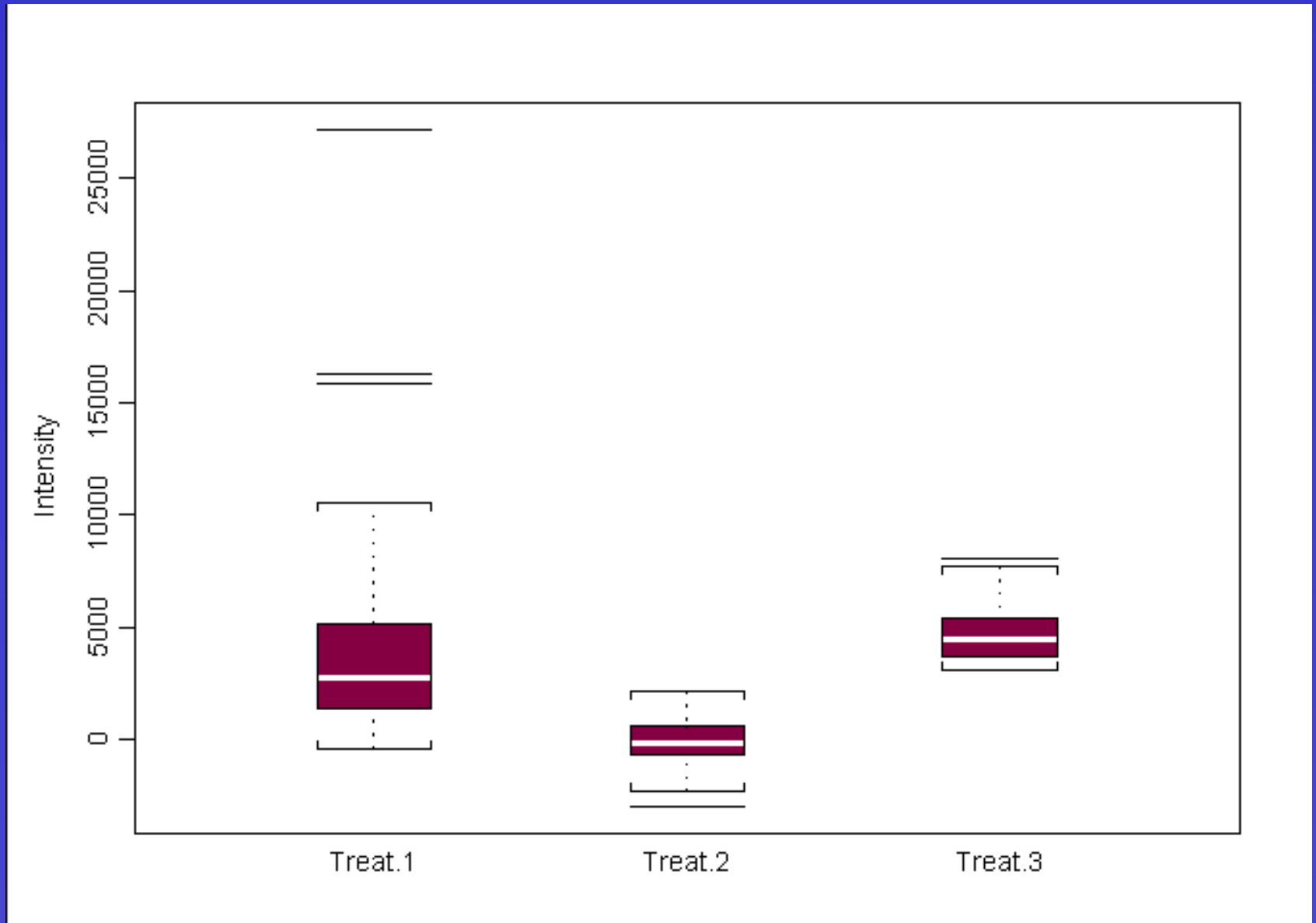
ALL (Sample 5)



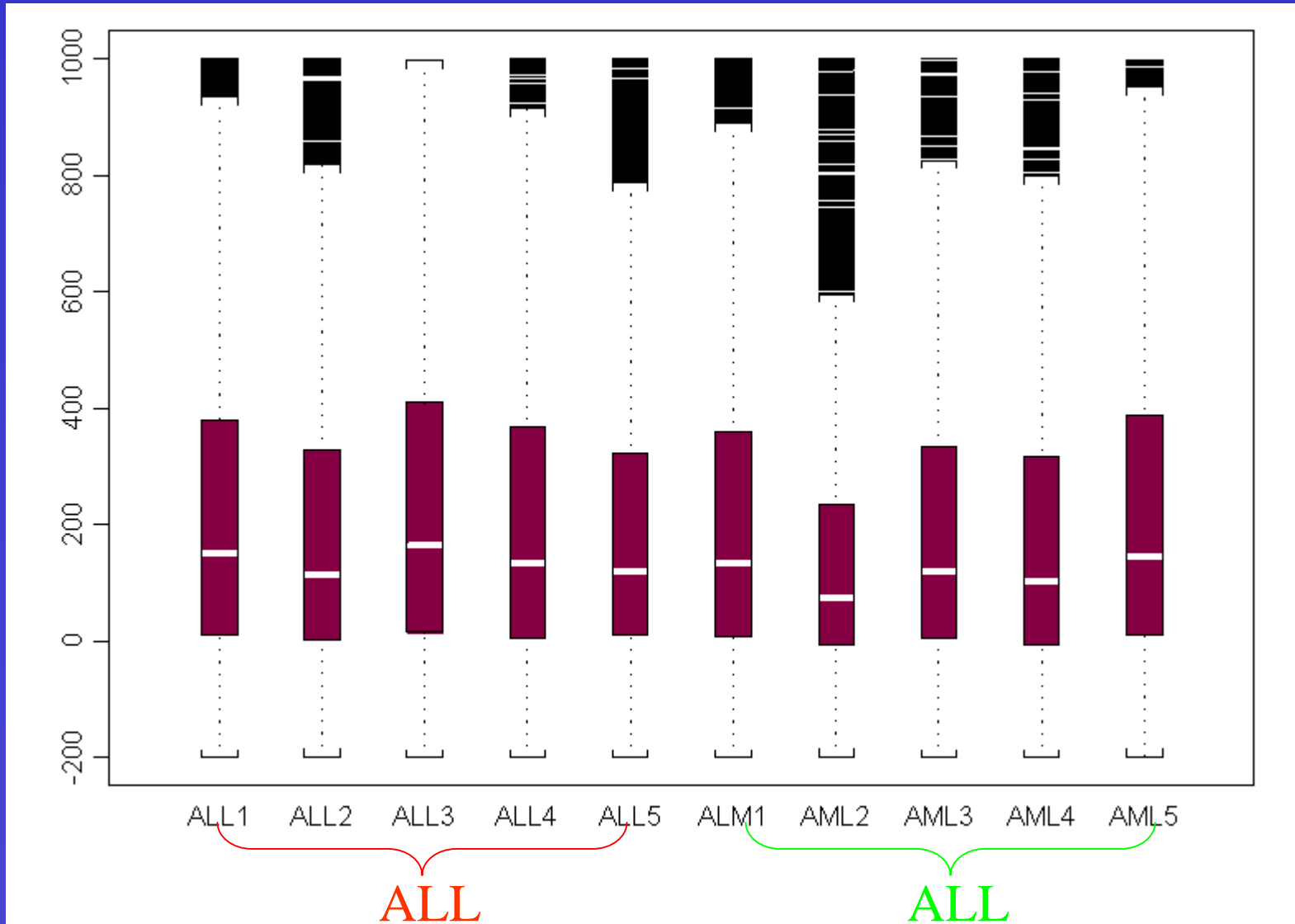
AML (Sample 5)



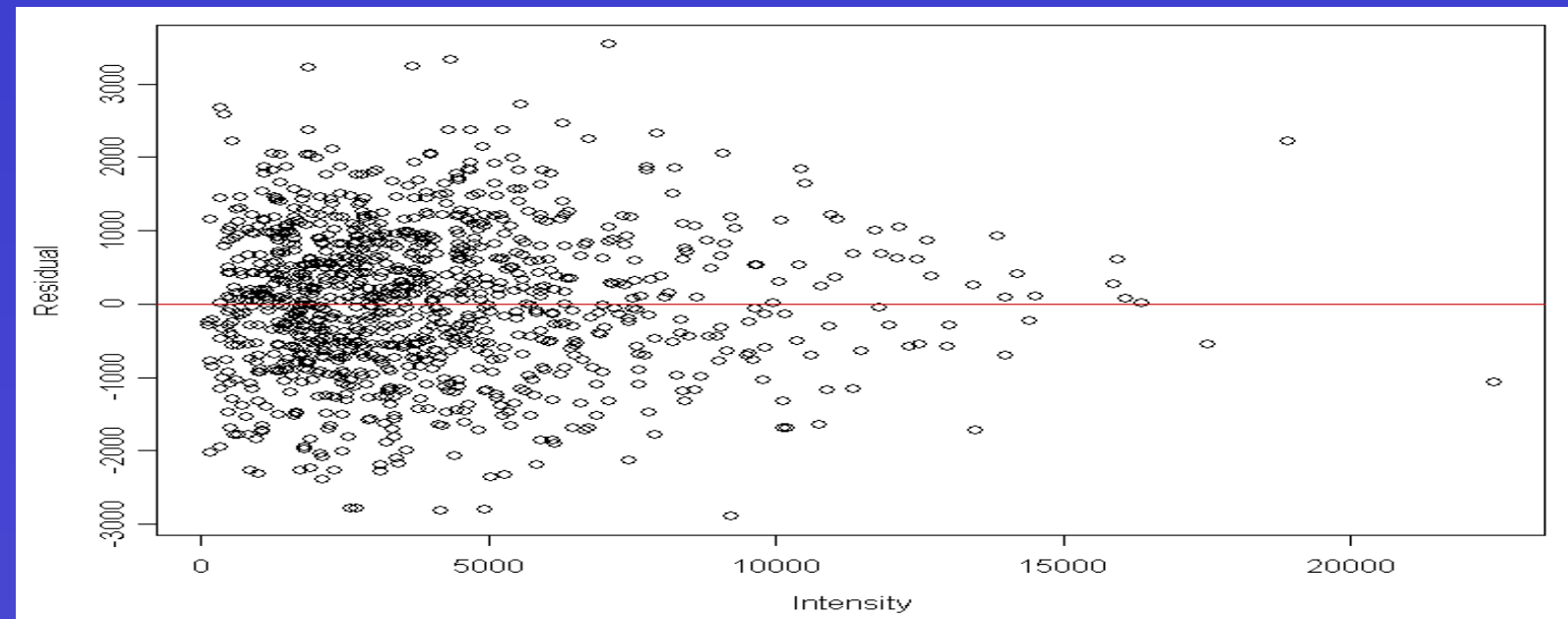
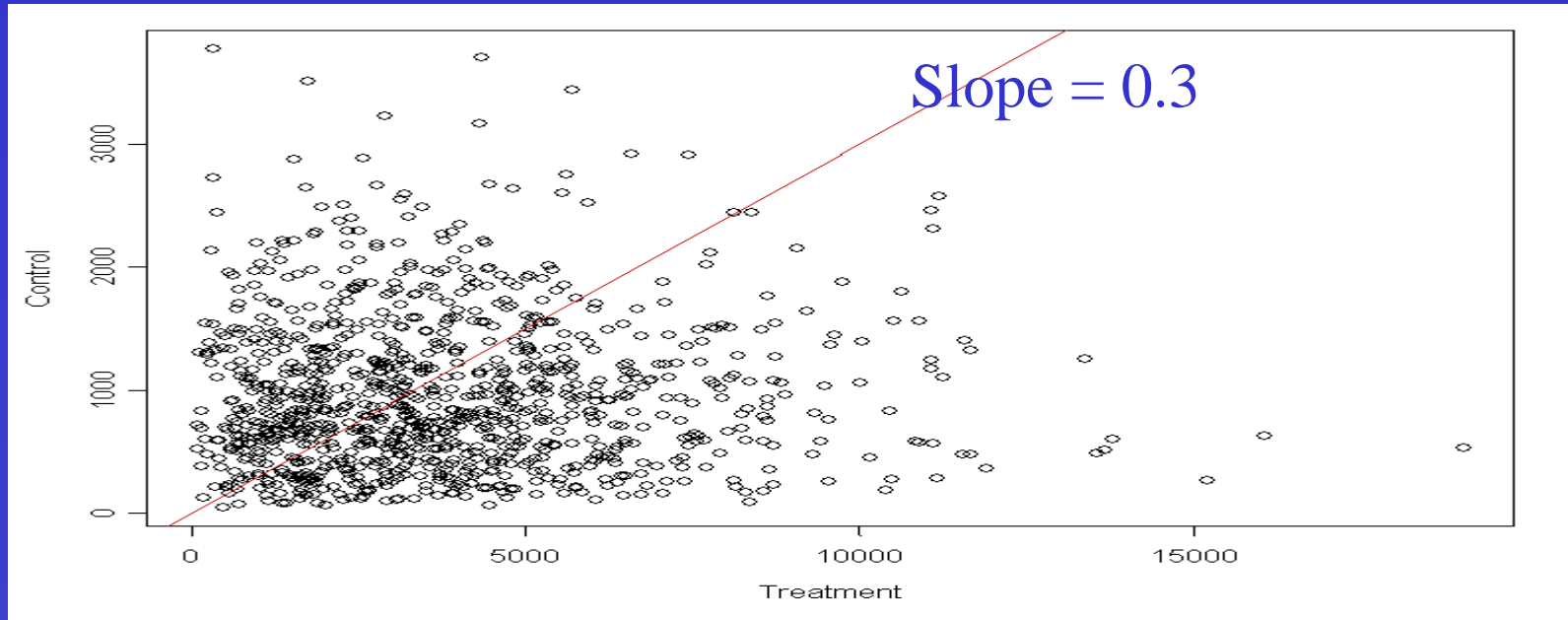
Box-plot



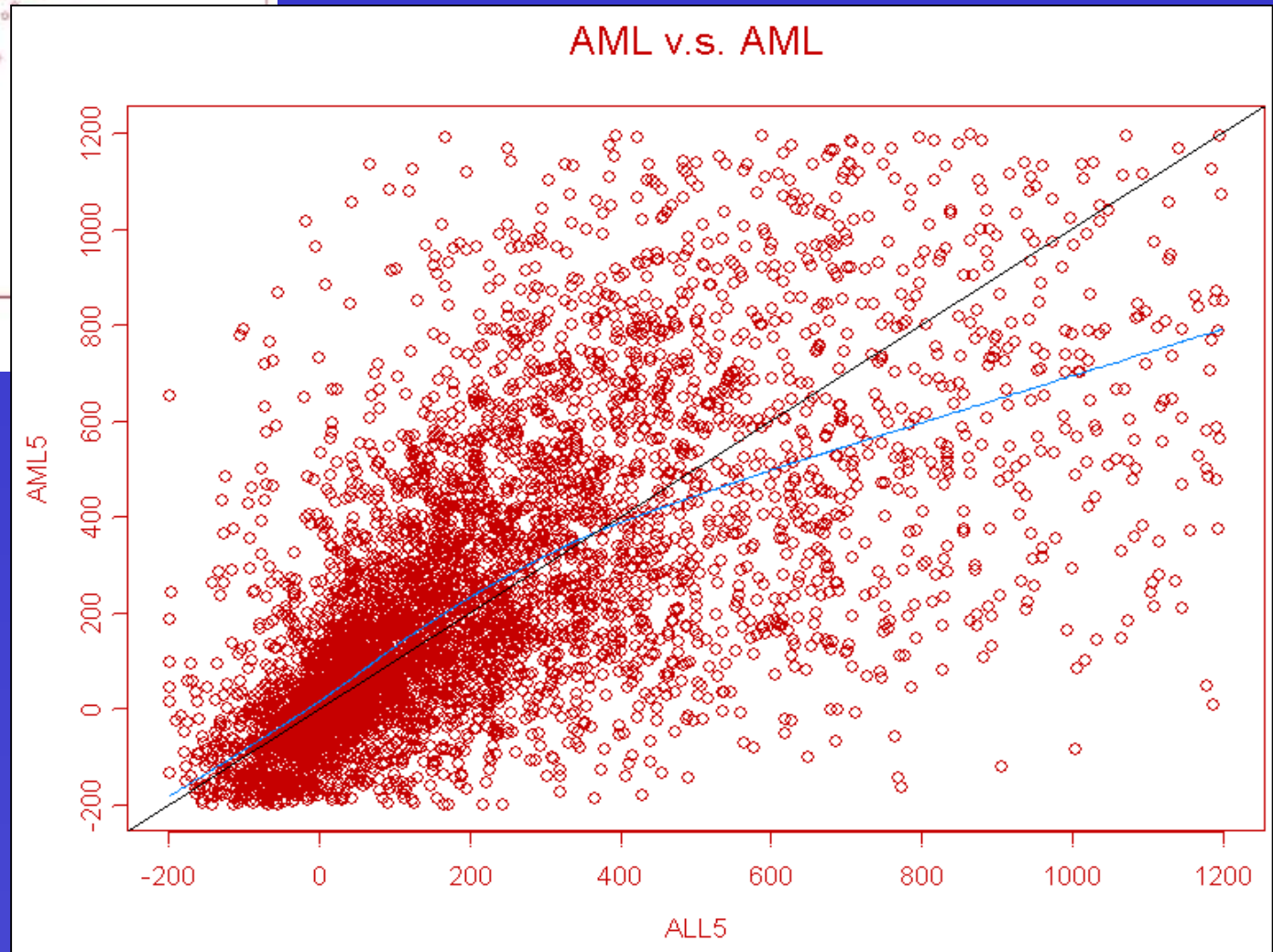
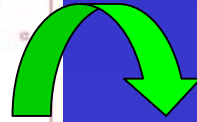
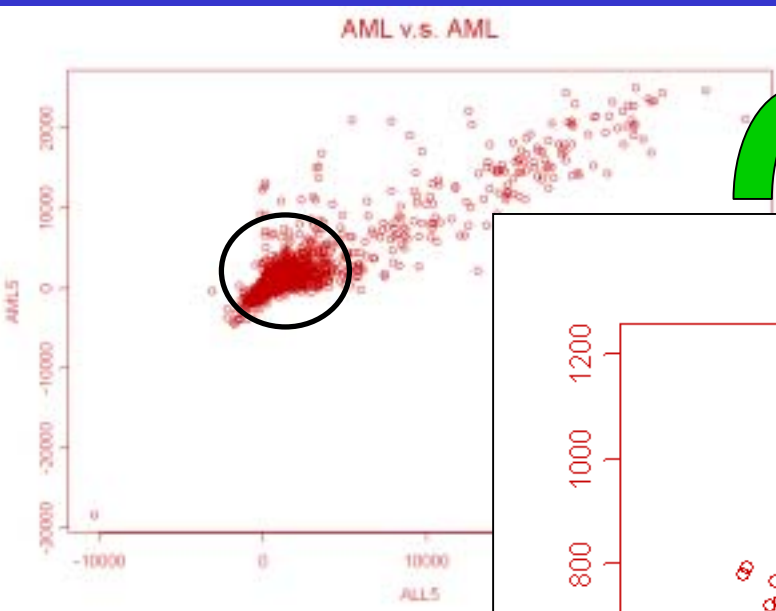
Example: ALL v.s. AML (Sample1~5)



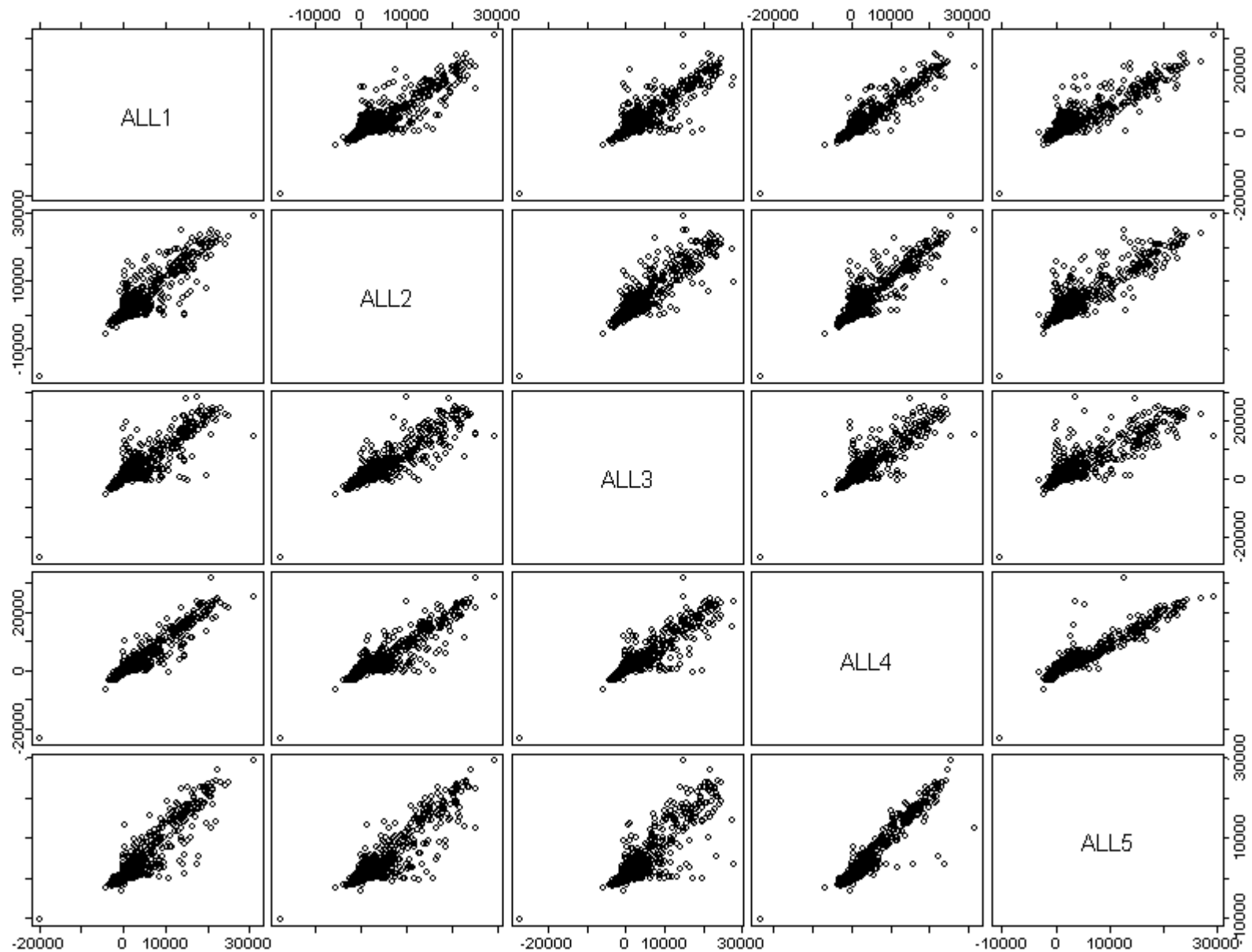
Scatter Plot



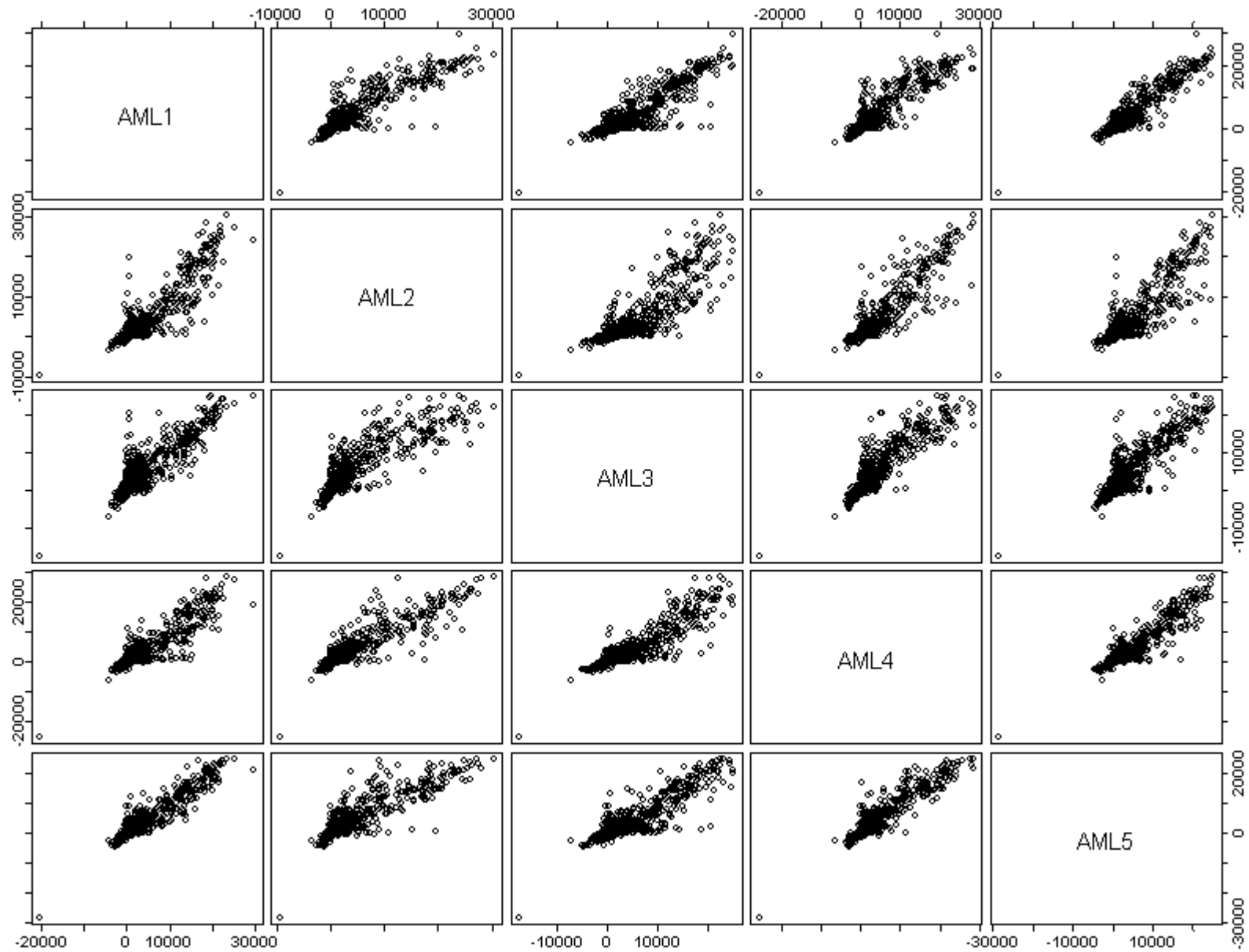
Example: ALL v.s. AML (Sample 5)



Example: Pairs (ALL, Sample1~5)



Example: Pairs (AML, Sample1~5)



Statistical Distributions

Discrete Type

- Bernoulli distribution
- Binomial distribution
- Hypergeometric distribution
- Multinomial distribution
- Poisson distribution
- ...

Continuous Type

- Normal distribution
- Gamma distribution
- Exponential distribution
- Chi-square distribution
- Student's t distribution
- F distribution
- ...

Bernoulli Distribution

An experiment that can have one(X) of **two** outcomes:

Success(S , $x=1$), Failure(F , $x=0$)  **Bernoulli experiment**

$$P(S) = P(X=1) = p, \quad P(F) = P(X=0) = 1 - p = q$$

Probability distribution function

$$f(x) = P(X = x) = \begin{cases} p^x (1-p)^{1-x}, & x = 0, 1 \\ 0 & , o.w. \end{cases}$$

$$X \sim Ber(p)$$

$$\mu = E(X) = p$$

$$\sigma^2 = Var(X) = E(X - \mu)^2 = E(X^2) - \mu^2 = p - p^2 = pq$$

Binomial Experiment

Repeated n times independent Bernoulli experiment

 Binomial experiment

i.e. a binomial experiment possesses the following properties:

- 1⁰ the experiment consists of a fixed number n of trials
- 2⁰ the result of each trial can be classified into one of two categories
- 3⁰ the probability p of a success remains constant for each trial
- 4⁰ each trial of the experiment is independent of the other trials

Binomial Distribution

Let r.v. X be the number of successes in the n trials of a Binomial experiment, then X is called the Binomial distribution, $X \sim \text{Bin}(n, p)$

The probability distribution function is

$$f(x) = P(X = x) = \begin{cases} C_x^n p^x (1-p)^{n-x}, & x = 0, 1, 2, \dots, n \\ 0 & , \text{ o.w.} \end{cases}$$

where
$$C_x^n = \frac{n!}{x!(n-x)!}$$

mean
$$\mu = E(X) = np$$

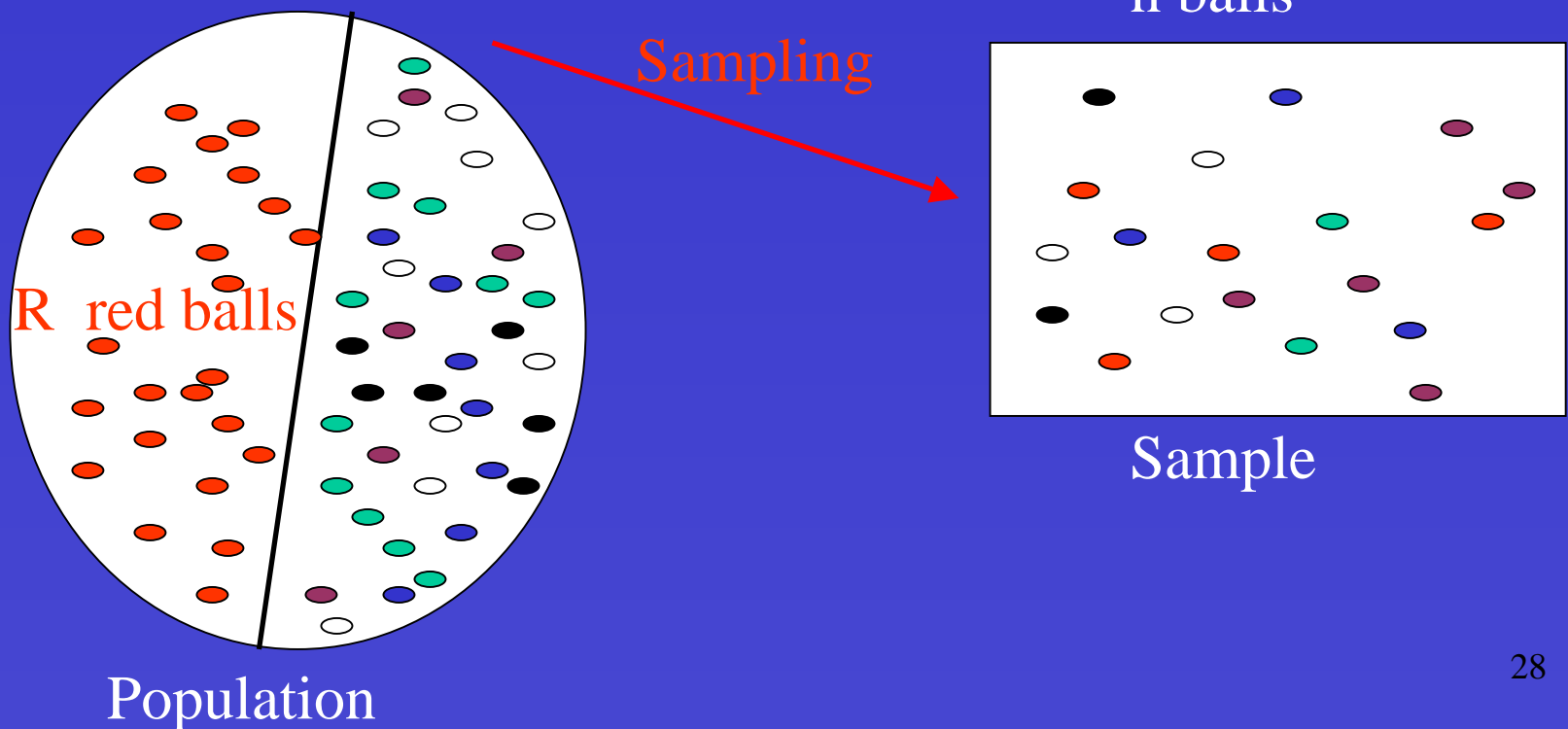
variance
$$\sigma^2 = \text{Var}(X) = E(X - \mu)^2 = E(X^2) - \mu^2 = np(1-p)$$

Hypergeometric Distribution

Sampling with replacement (**WR**) \ggg Binomial

v.s.

Sampling without replacement (**WTR**) \ggg Hyp.



Hypergeometric Distribution(cont.)

Let X be the number of red balls in the sample, then the distribution of X is the hypergeometric distribution, $X \sim Hyp(n, R, N)$

The probability distribution function is

$$f(x) = P(X = x) = \begin{cases} \frac{C_x^R C_{n-x}^{N-R}}{C_n^N}, & \max\{0, n - N + R\} \leq x \leq \min\{n, R\} \\ 0 & , \text{ o.w.} \end{cases}$$

mean $\mu = E(X) = n \frac{R}{N}$ variance $\sigma^2 = Var(X) = n \frac{R}{N} \left(1 - \frac{R}{N}\right) \frac{N-n}{N-1}$

Theorem :

If $X \sim Hyp(n, R, N)$, then for each value $x = 0, 1, 2, \dots, n$,

and as $N \rightarrow \infty$, $R \rightarrow \infty$, with $\frac{R}{N} \rightarrow p$ (a positive constant),

$$\lim_{N \rightarrow \infty} \frac{C_x^R C_{n-x}^{N-R}}{C_n^N} = C_x^n p^x (1-p)^{n-x}$$

Multinomial Distribution

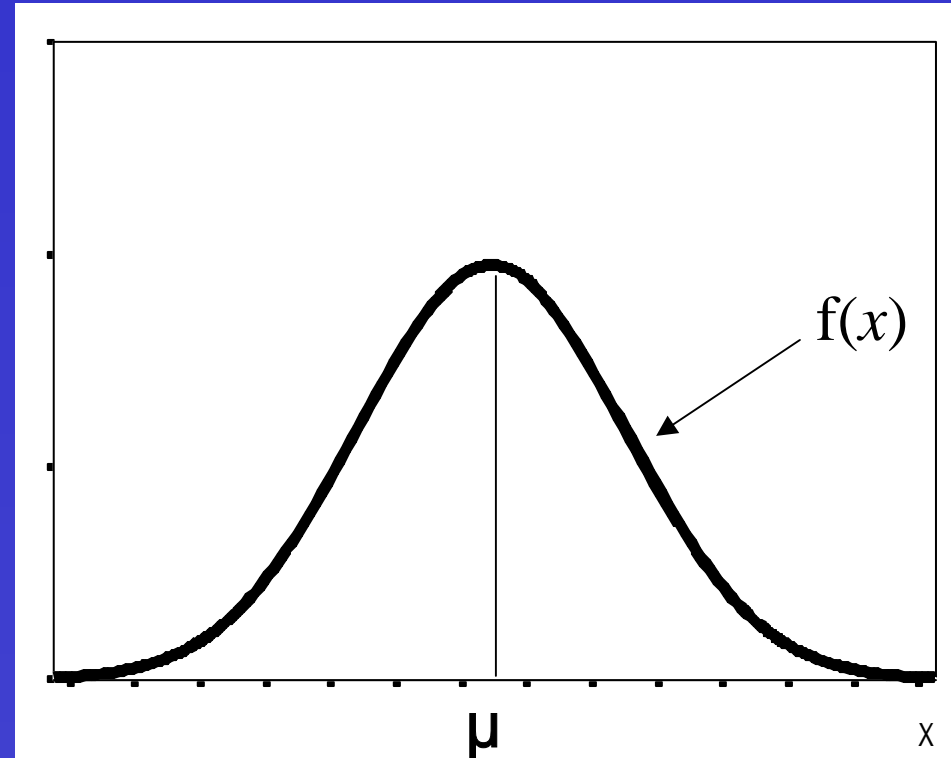
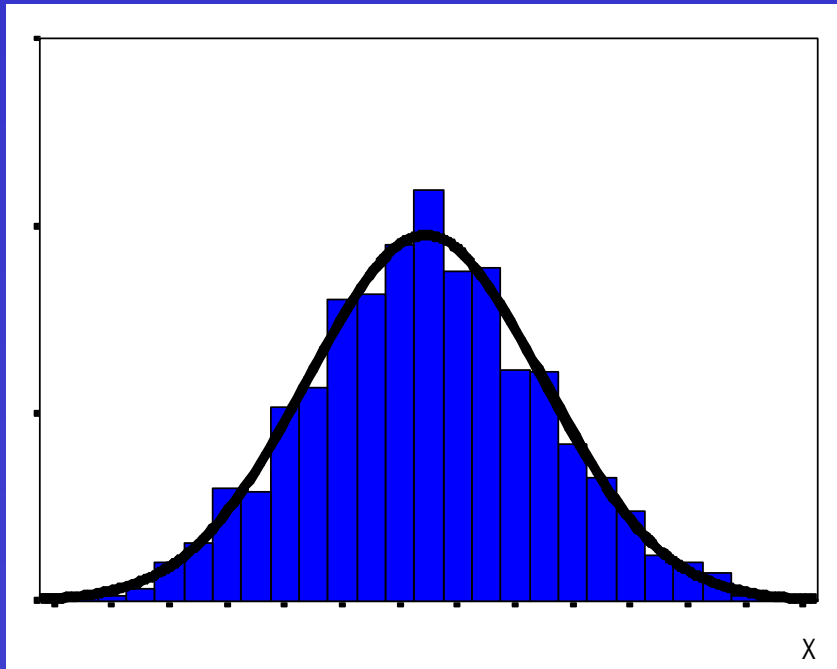
If each trial has several different outcomes, label the different possible types resulting from each trial by i where $i = 1, 2, \dots, k$, the probability of each type at each trial is p_i , and the count of each of the types in a sample of size n as X_i , then the probability of

$X = (X_1, X_2, \dots, X_k)$ is

$$\begin{aligned} f(x_1, x_2, \dots, x_k) &= P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) \\ &= \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k} \\ &= \frac{n!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k p_i^{x_i}, \quad \sum_{i=1}^k x_i = n, \sum_{i=1}^k p_i = 1 \end{aligned}$$

$$X \sim \text{MULT}(n, p_1, p_2, \dots, p_k)$$

Normal Distribution

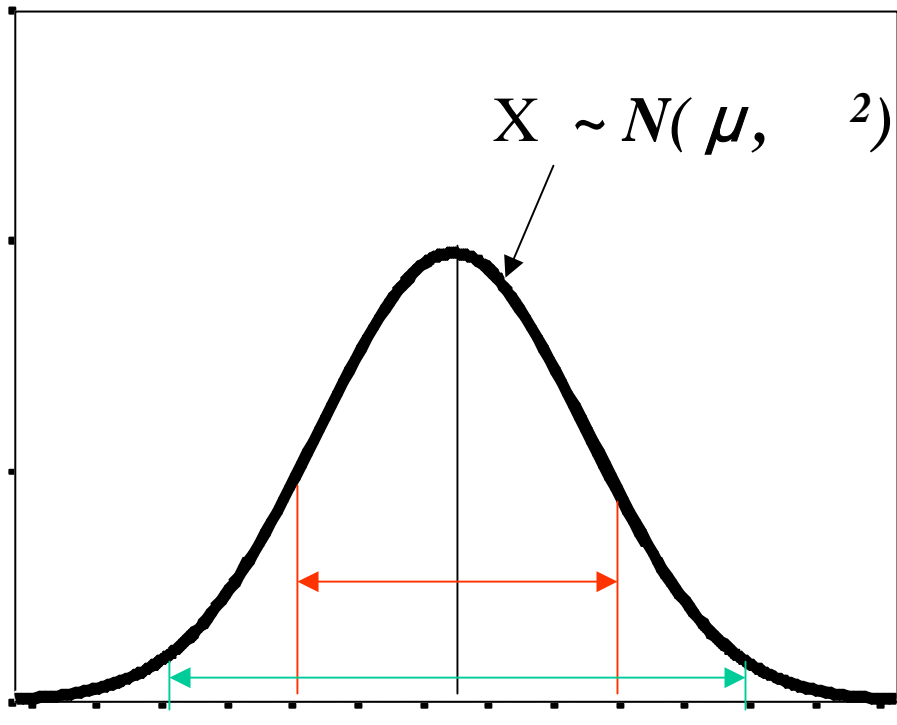


r.v. $X \sim N(\mu, \sigma^2)$

the pdf for X is $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$, $-\infty < \mu < \infty$, $\sigma > 0$

$E(X) = \mu$, $Var(X) = \sigma^2$

Normal Distribution(cont.)



$\mu - 2$ μ $\mu + 2$

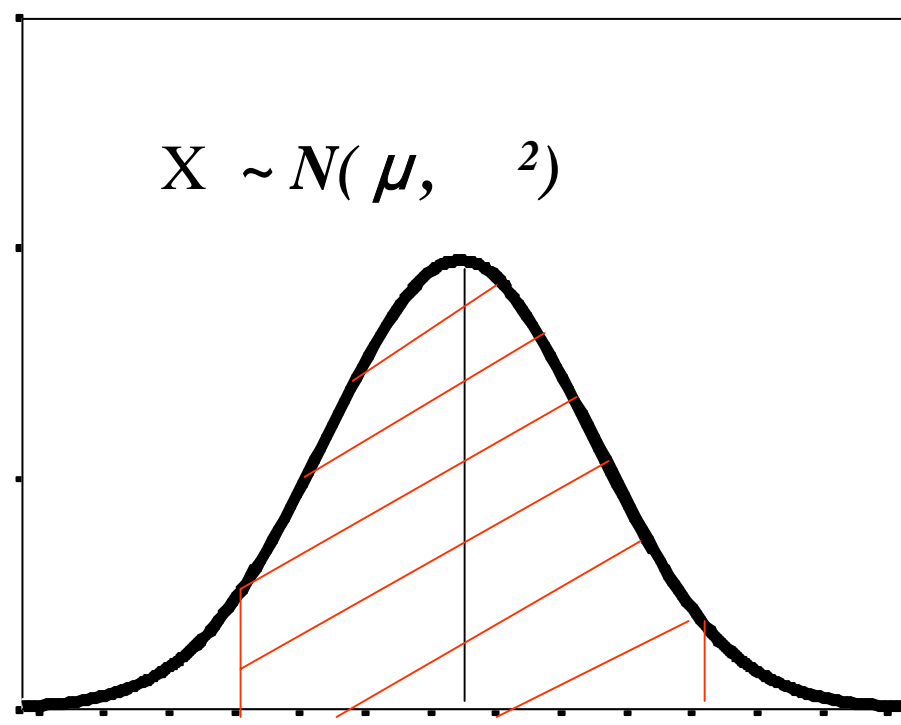
$\mu - 2$

$\mu + 2$

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.683$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.954$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.997$$



a μ b

$$P(a \leq X \leq b) = F(b) - F(a)$$

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

Normal Distribution(cont.)

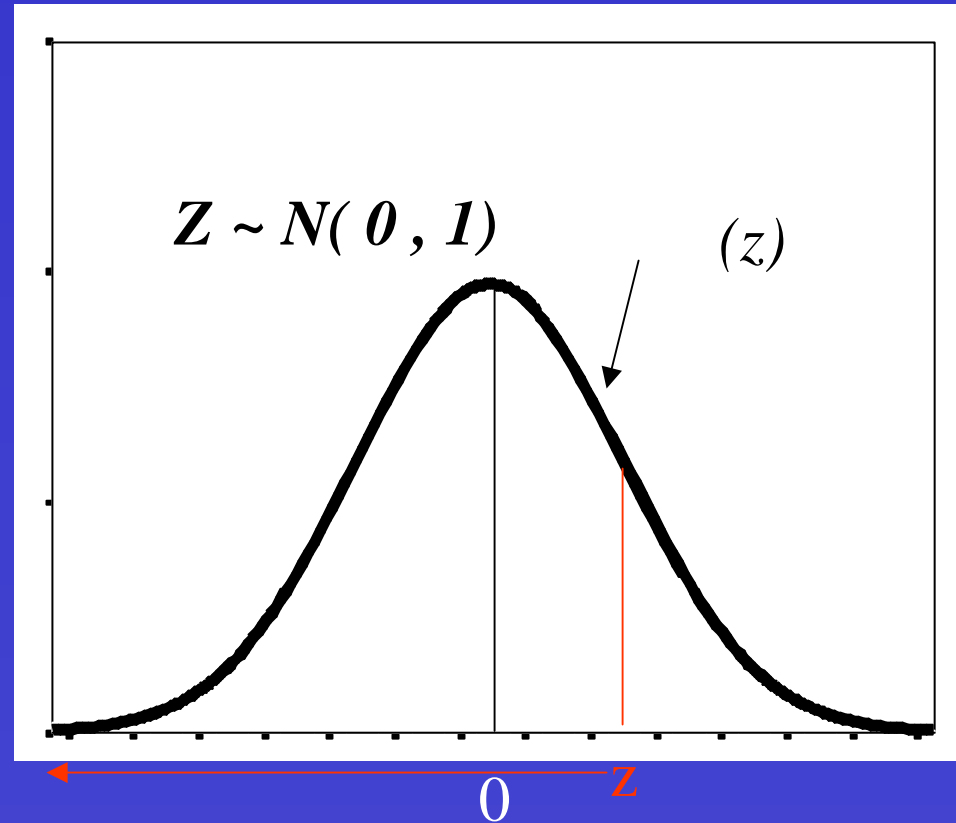
$X \sim N(\mu, \sigma^2)$ **normalized** $Z \sim N(0, 1)$

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

the pdf for X is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}},$$

$$-\infty < z < \infty$$



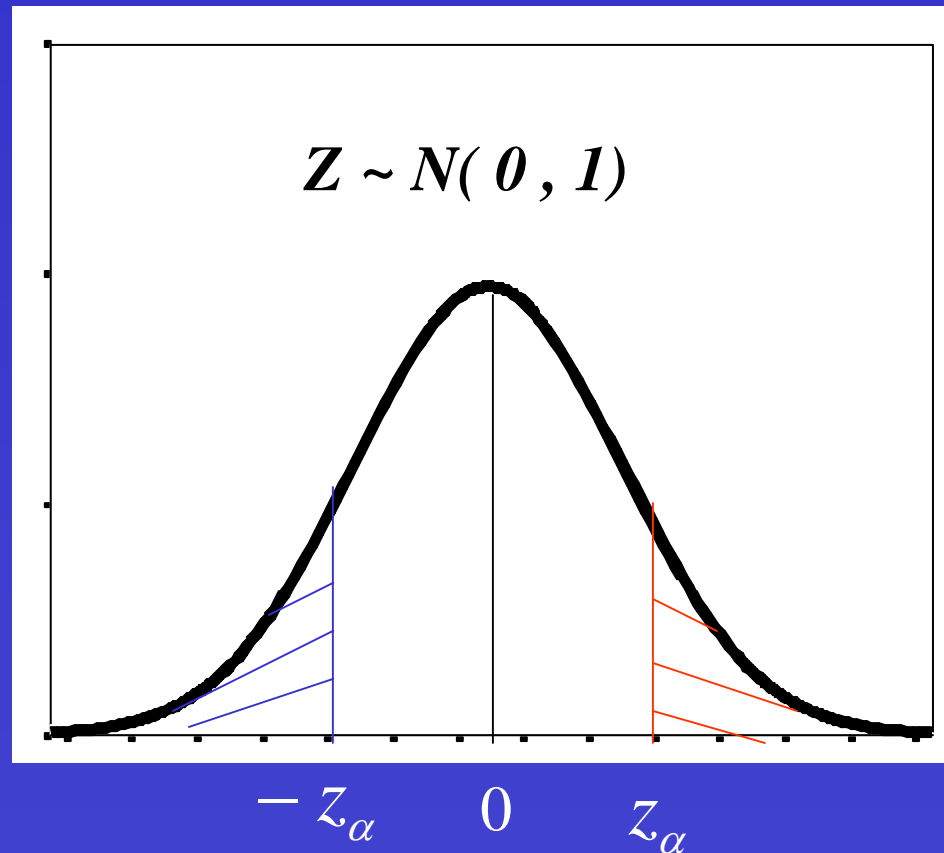
$$P(Z \leq z) = \Phi(z) = \int_{-\infty}^z \phi(z) dz = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = ?? \quad (\text{查表})$$

Normal Distribution(cont.)

Standard Normal Cumulative Distribution Function (z)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857

Normal Distribution(cont.)



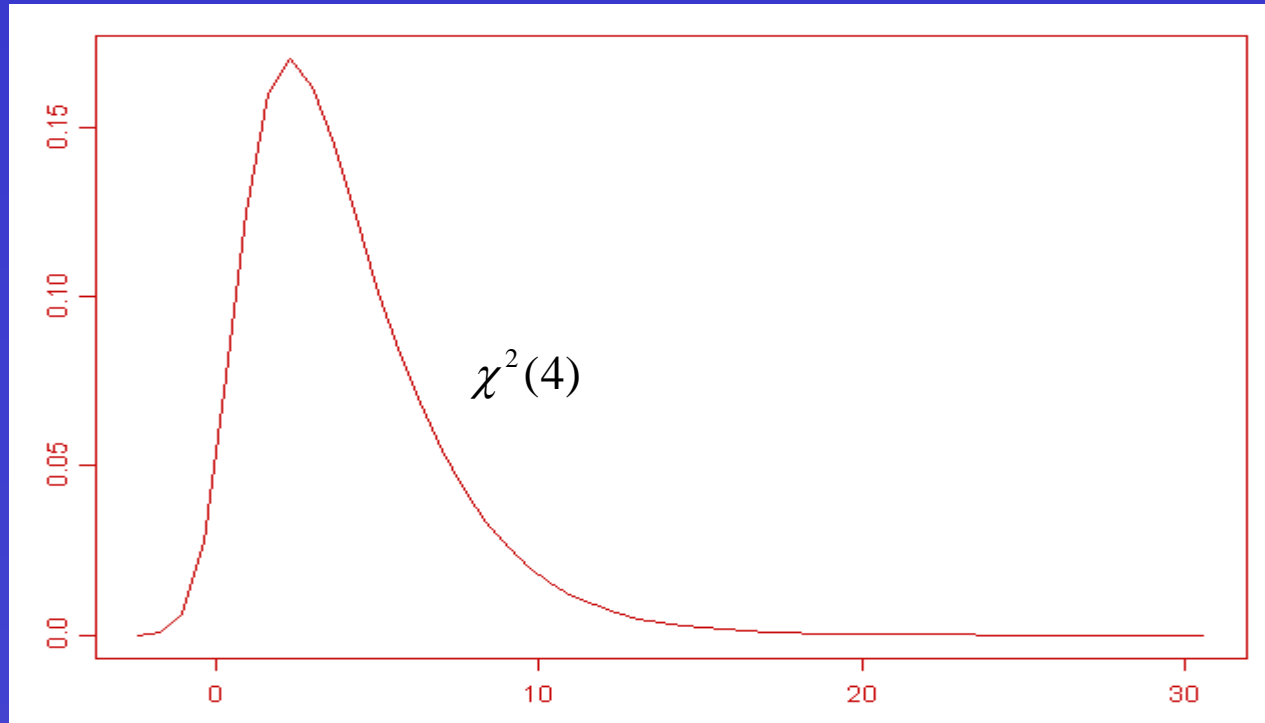
$$P(Z \geq z_\alpha) = P(Z \leq -z_\alpha) = \alpha$$

$$\Phi(z_\alpha) = 1 - \Phi(-z_\alpha) \Rightarrow \Phi(z_\alpha) + \Phi(-z_\alpha) = 1$$

例: $z_{0.025} = 1.96$, $z_{0.05} = 1.645$

Chi-square Distribution

$$r.v. X \sim \chi^2(\nu), \text{ pdf } f(x) = \frac{1}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}, x > 0$$
$$E(X) = \nu, \quad \text{Var}(X) = 2\nu$$



Remark: Random sample $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$

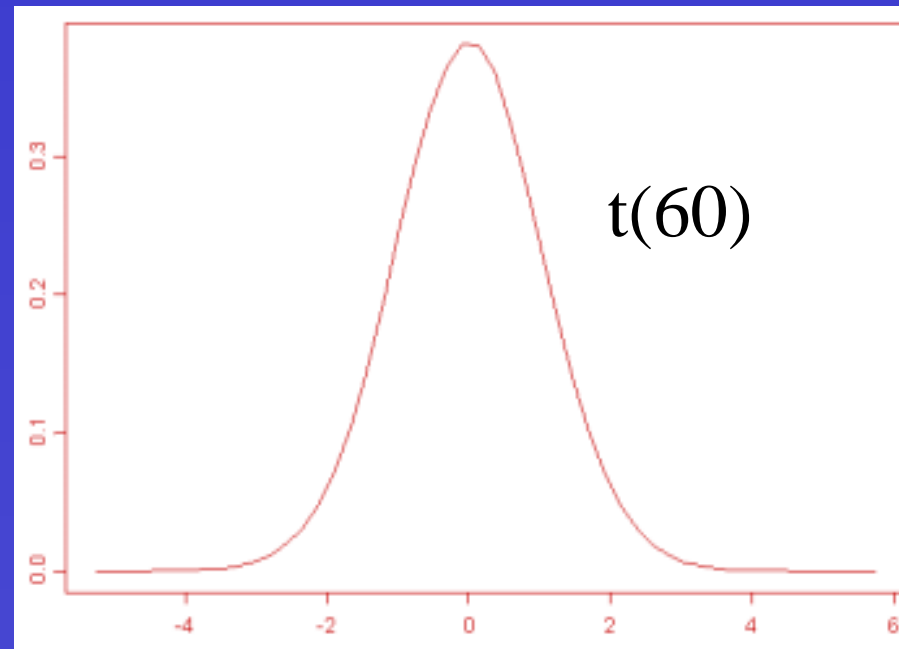
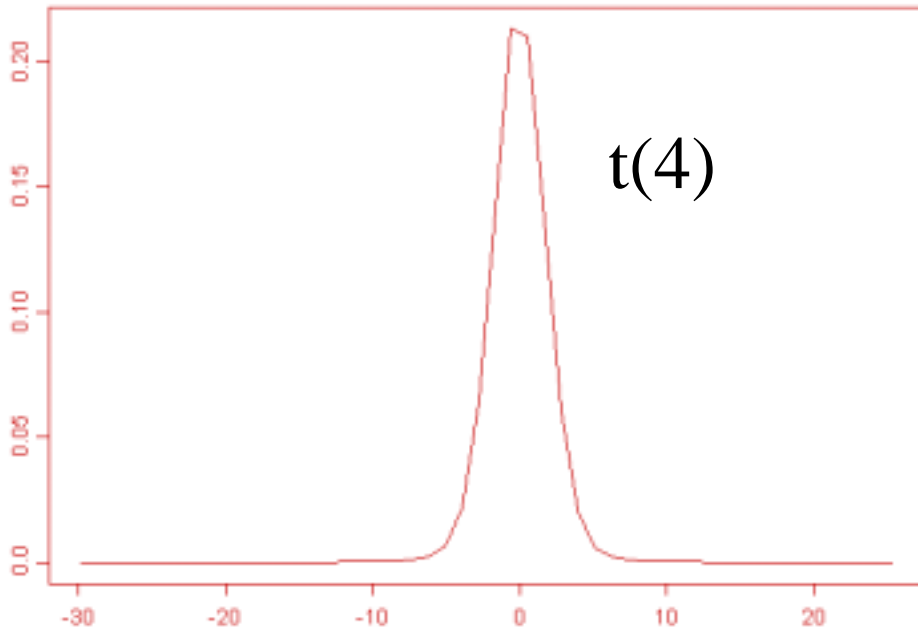
$$\Rightarrow \chi^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Student's t Distribution

if $Z \sim N(0,1)$ and $V \sim \chi^2(\nu)$ are independent then $T = \frac{Z}{\sqrt{V/\nu}} \sim t(\nu)$

$$\text{pdf } f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, x \in R$$

$$E(X) = 0, \quad \text{Var}(X) = \frac{\nu}{\nu-2}$$



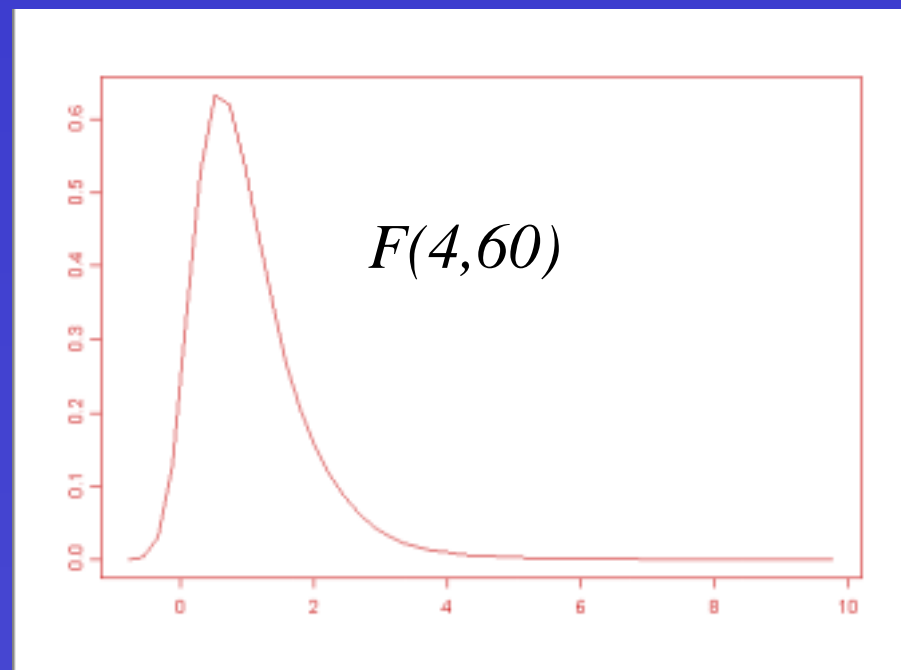
F Distribution

if $V_1 \sim \chi^2(v_1)$ and $V_2 \sim \chi^2(v_2)$ are independent then $X = \frac{V_1/v_1}{V_2/v_2} \sim F(v_1, v_2)$

$$\text{pdf } f(x) = \frac{\Gamma\left(\frac{v_1 + v_2}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)} \left(\frac{v_1}{v_2}\right)^{v_1/2} x^{(v_1/2)-1} \left(1 + \frac{v_1}{v_2}x\right)^{-\frac{v_1+v_2}{2}}, \quad x > 0$$

$$E(X) = \frac{v_2}{v_2 - 2}, \quad 2 < v_2$$

$$\text{Var}(X) = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)}, \quad 4 < v_2$$



Statistical Inference

Estimation of parameter

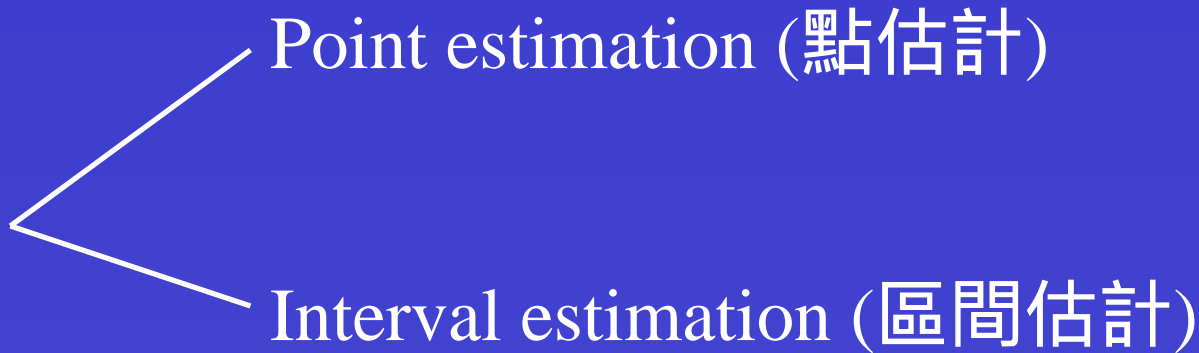
Testing of statistical hypothesis

Terminologies

- **Population** : the set of measurements corresponding to the entire collection of units about which information is sought
- **Sample** : the set of measurements that actually collected in the course of an investigation
- **Parameter** : a numerical feature of a population
- **Statistic** : a numerical valued function of the sample observations

Estimation

Estimation (估計) : to assign an appropriate value (or interval) for a parameter based on observed data from the population



Point Estimation

$$\hat{\theta} \text{ (Statistic)} \xrightarrow{\text{Est.}} \theta \text{ (Parameter)}$$

e.g.

$$\text{Sample mean } \bar{X} \xrightarrow{\text{Est.}} \text{Population mean } \mu$$

$$\text{Sample variance } S^2 \xrightarrow{\text{Est.}} \text{Population variance } \sigma^2$$

$$\text{Sample proportion } \hat{p} = \frac{x}{n} \xrightarrow{\text{Est.}} \text{Population proportion } p$$

Interval Estimation

Sample mean \bar{X} $\xrightarrow{\text{estimate}}$ Population mean μ (accuracy ?)

Interval estimation : using an interval to estimate unknown population parameter

Definition: Confidence Interval(C.I.)

An interval (l, u) is called a $100\gamma\%$ C.I. for parameter θ , if

$$P(l \leq \theta \leq u) = \gamma, \quad 0 < \gamma < 1, \quad \text{and}$$

$l = l(X_1, X_2, \dots, X_n), u = u(X_1, X_2, \dots, X_n)$ are statistics.

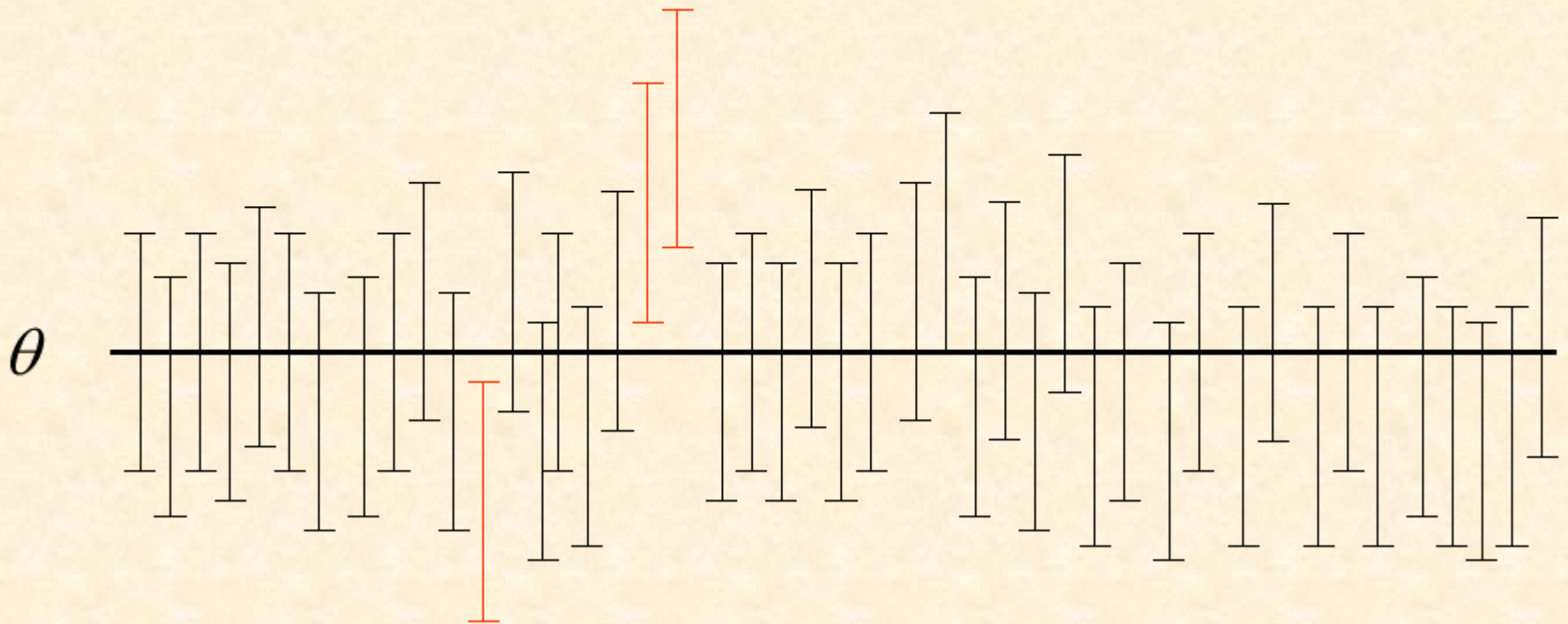
Interval Estimation(cont.)

Remark: C.I.

- (1) the observed values $l = l(x_1, x_2, \dots, x_n)$ and $u = u(x_1, x_2, \dots, x_n)$ are called lower and upper confidence limits, respectively
- (2) γ is called the confidence coefficient, confidence level or degree of confidence, e.g. 0.90, 0.95, 0.99
- (3) If such interval estimates are computed from many different samples, then in the long run we would expect about $100\gamma\%$ of the intervals to include the true value of unknown parameter θ .

Interval Estimation(cont.)

Example: C.I.



Example---Estimation

The following data are the signal of a gene expression on the microarray experiment :

1014, 1400, 2116, 1565, 1560, 1000, 2130, 1570, 1730.

Point estimation:

Sample mean	1565
Sample median	1565
Sample variance	162519

Interval estimation:

90%	[1255 , 1875]
95%	[1195 , 1935]
99%	[1151 , 1979]

Introduction---Testing

- Is a new drug effective ? *e.g.* $p \geq 80\%$?
- Is the mean lifetime of a component at least some specified amount ?
e.g. $\mu \geq 2 \text{ years}$?
- Does a lot of manufactured items contain an excessive number of defectives ? *e.g.* $p_D \geq 3\%$?
- In a microarray experiment for a gene, can it be concluded that the mean intensity in treatment exceeds that in control ? *e.g.* $\mu_T \geq \mu_C$
- the mean gene expression of a treatment group is at least some specified amount *e.g.* $\mu \geq \mu_0$
- the variance of the gene expression of a treatment group is not great than some specified amount *e.g.* $\sigma^2 \leq \sigma_0^2$
- many biological decisions require a determination of whether the means (variance) of two (or some) treatments differ

Hypothesis Testing

Hypothesis testing : the process of trying to decide, on the basis of experimental evidence, the truth or falsity of the hypothesis

Null hypothesis H_0 : nullifies the research hypothesis

Alternative hypothesis H_a (or H_1) : the claim or the research hypothesis that we wish to establish

e.g. 1^o $H_0 : \theta \geq \theta_0$ v.s. $H_a : \theta < \theta_0$

2^o $H_0 : \theta \leq \theta_0$ v.s. $H_a : \theta > \theta_0$

3^o $H_0 : \theta = \theta_0$ v.s. $H_a : \theta \neq \theta_0$

Decision Rule

Either to **reject** null hypothesis and conclude that alternative hypothesis is substantiated
Or to **retain** null hypothesis and conclude that alternative hypothesis fails to be substantiated

- **Test statistic** : the sample statistic upon which we base our decision to either reject or not reject H_0
- **Rejection region(or Critical region)** : the subset of the sample space that corresponds to reject H_0

e.g. **Test** $H_0 : \mu \geq \mu_0$ v.s. $H_a : \mu < \mu_0$

Test statistic : Sample mean \bar{X}

Rejection region : $\bar{X} < C$

Decision Errors

- Decision errors :

Type I error --- reject H_0 if H_0 is true

Type II error --- do not reject H_0 if H_0 is false

- Error probabilities :

$$\alpha = P(\text{Type I error}) = P(\text{reject } H_0 | H_0)$$

$$\beta = P(\text{Type II error}) = P(\text{do not reject } H_0 | H_a)$$

	H_0 is true	H_0 is false
Reject H_0	Type I error, α	Correct decision
Do not reject H_0	Correct decision	Type II error, β

Significance Level & P-value

- the significance level of the test is

$\alpha = P(\text{Type I error})$, for a simple null hypothesis

$\text{Max } P(\text{Type I error})$, for a composite null hypothesis

➔ to determine a rejection region

- P-value :

the smallest size α at which H_0 can be rejected,
based on **the observed value of the test statistic**

- the P -value serves as a measure of the strength of evidence against null hypothesis H_0
- a small P -value means that the null hypothesis H_0 is strongly rejected

Example---Testing

The following data are the signal of a gene expression on the microarray experiment, 1014, 1400, 2116, 1565, 1560, 1000, 2130, 1570, 1730.

Assume that the gene expression is a normal distribution and its population variance is 10000.

Test the hypothesis that the mean signal of the gene is less than 1500.

Hypothesis $H_0 : \mu \geq 1500$ v.s. $H_1 : \mu < 1500$

Test statistic \bar{X}

Rejection region $\bar{X} < C$

P-value $P(\bar{X} < 1565 | H_0) \approx 0.026$

Example---Gene Expressions

Control: 20741, 20630, 15822, 20999, 17358, 22600, 25056
Treatment: 20780, 19553, 25766, 27682, 24955, 25946, 24995

Test $H_0 : \mu_{Control} = \mu_{Treatment}$

Results

Method	Test-Statistic	P-value	d.f.
t-Test(equal variance)	-2.3437	0.0371	12
t-Test(unequal variance)	-2.3437	0.0372	11.974
Paired t-Test	-2.3002	0.0611	6
Wilcoxon rank-sum test	38.0	0.0728	
Wilcoxon sign-rank test	5.0	0.1562	

Comparing Two Treatments

Test variance $\sigma_1^2 = \sigma_2^2 \longrightarrow F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$

Compare the means

parametric

T-test ,
Paired T-test

nonparametric

Wilcoxon rank sum test ,
Wilcoxon sign rank test

F-test

- Testing for difference between two variances

If we have two samples of measurements, each sample taken at random from a normal population, we might ask if the variances of the two populations are equal.

Test $\sigma_1^2 = \sigma_2^2$

$$\text{Test statistic } F = \frac{\frac{(n_1 - 1)S_1^2}{\sigma_1^2}}{\frac{(n_2 - 1)S_2^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2} \frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1 - 1, n_2 - 1)$$

- Analysis of variance(ANOVA)

to test for differences among several population means

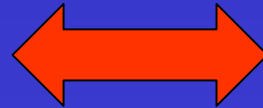
Cluster v.s. Classification

Two main types of microarray experiments:

- sample are not categorized, but gene expression levels are followed with time
 - to find genes whose expression levels vary together
 - **Cluster**
- samples are labelled as either normal or diseased tissues
 - to find genes whose expression levels can distinguish different labels
 - **Classification**

Statistical Pattern Recognition

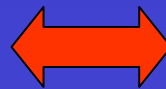
Clustering
(unsupervised)



Discrimination
(supervised)

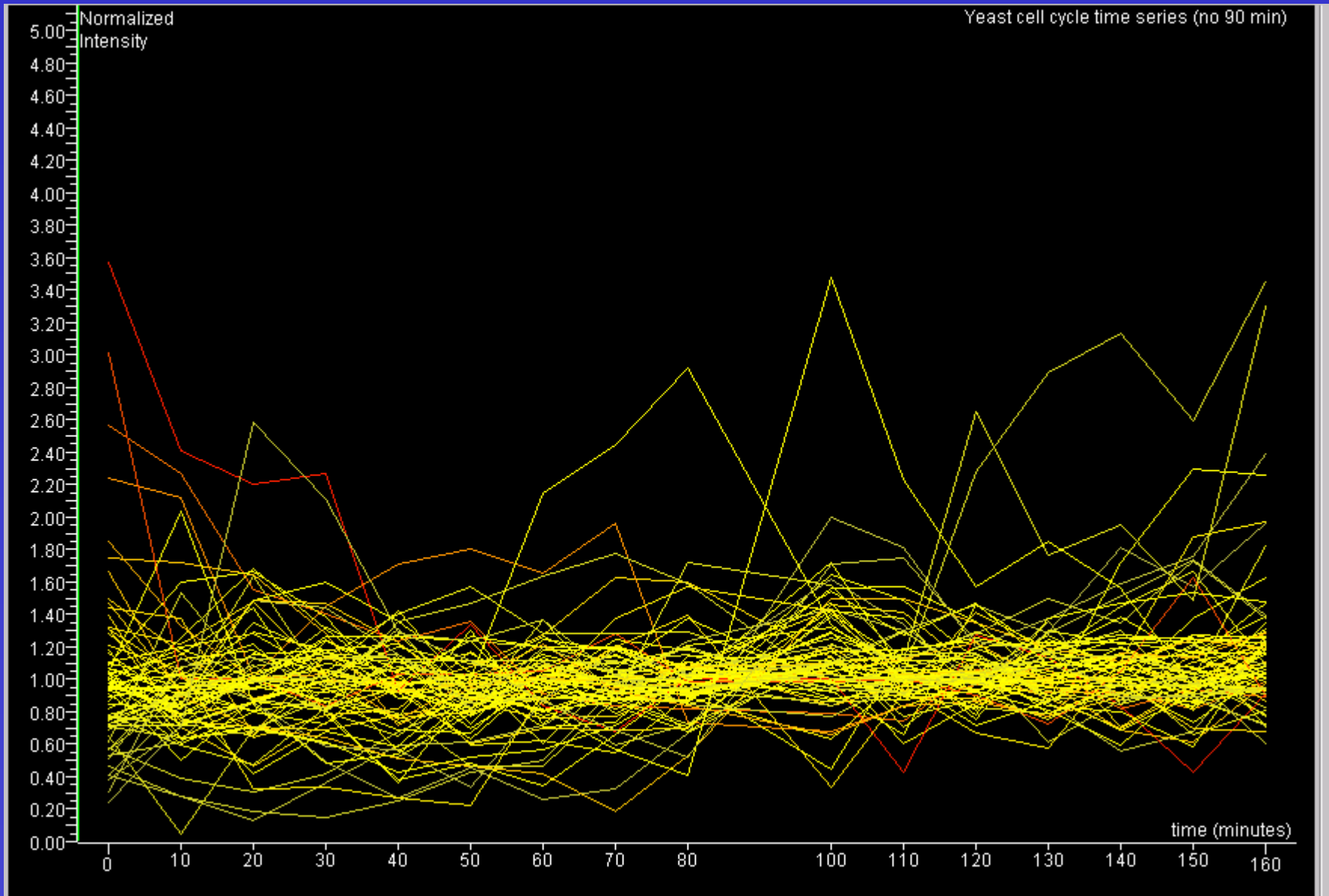


Nonparametric method



Parametric method

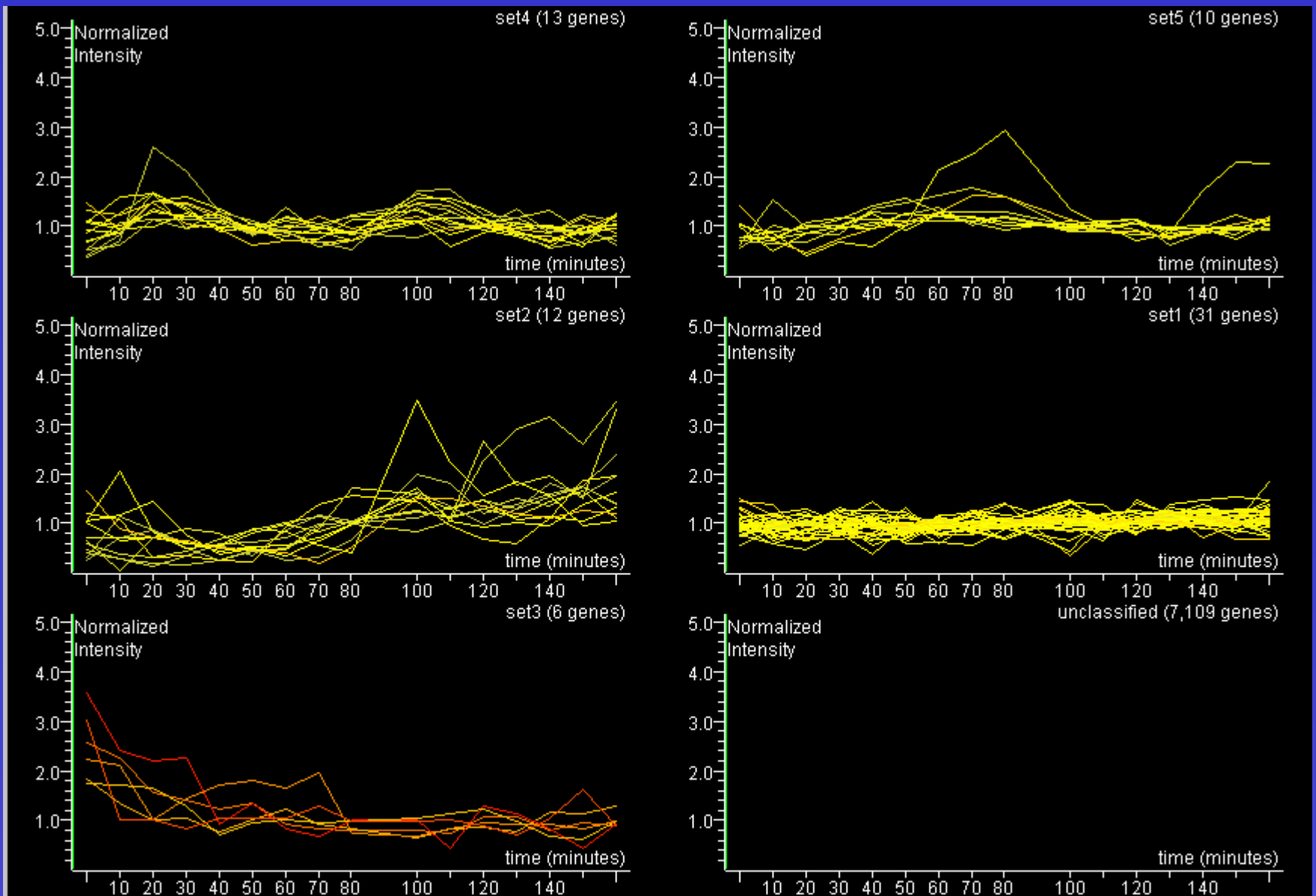
Time Course Data



time 0 minutes

Animate

Clustering



time 0 minutes

Animate

Cluster Analysis

Cluster analysis is the art of finding groups in data.
(Kaufman & Rousseeuw, 1990)

- Objective of clustering analysis :
 - one wants to form groups in such that
 1. **homogeneous** ---
objects in the same group are similar to each other
 2. **well separated** ---
objects in different groups are dissimilar as possible

Recent Methods---Cluster

- Partition methods : K-means
 - Partitioning around medoids (pam)
 - Fuzzy clustering
- Hierarchical methods : Agglomerative
 - Divisive
- Graph theoretic method : Self-organizing maps(SOM)

Discrimination v.s. Classification

Discrimination or **Separation** :

Separating distinct sets of genes

Classification, Allocation, or Prediction

Allocating new objects to previously defined groups

Classification or allocation rules are usually developed from learning(or training) sample

Multivariate techniques

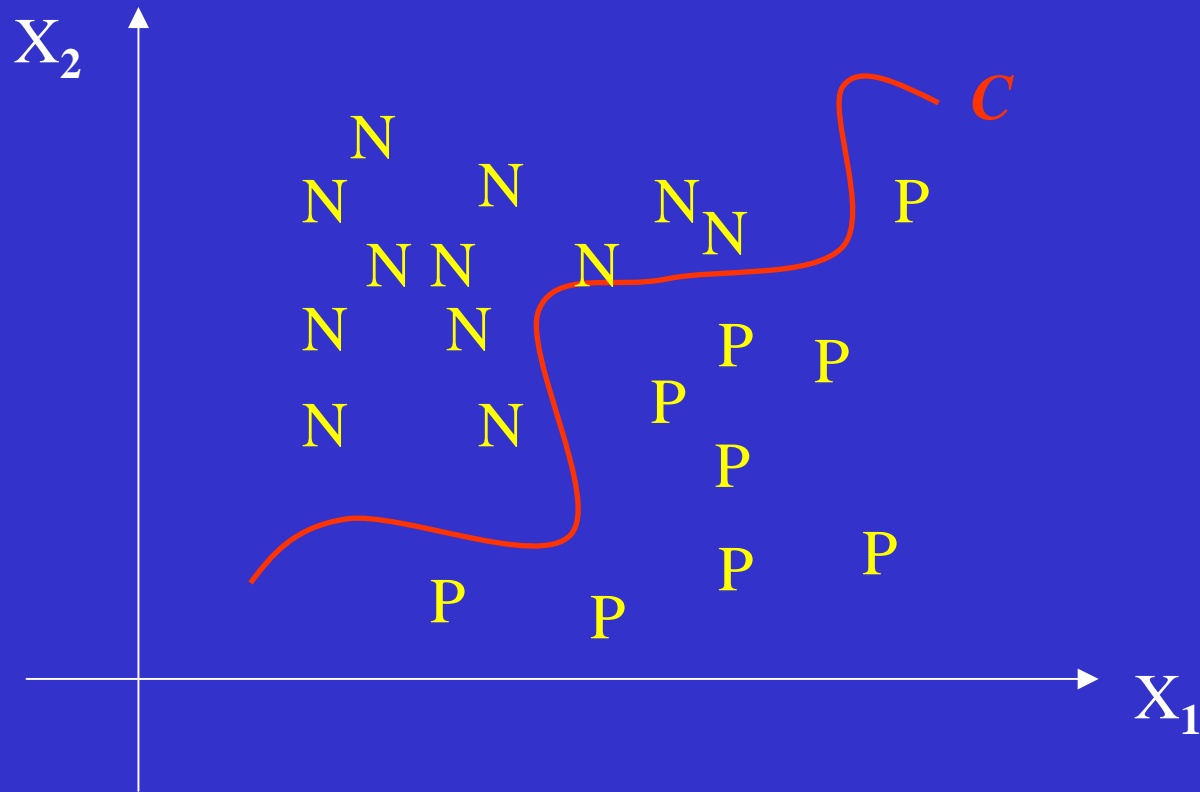
Require

A set of feature variables

- it is important for selecting better feature variables
- it depends on the master's insight to the nature of the subject

Given n objects, how can one assign each object into K known classes with **min. error (or regret)**

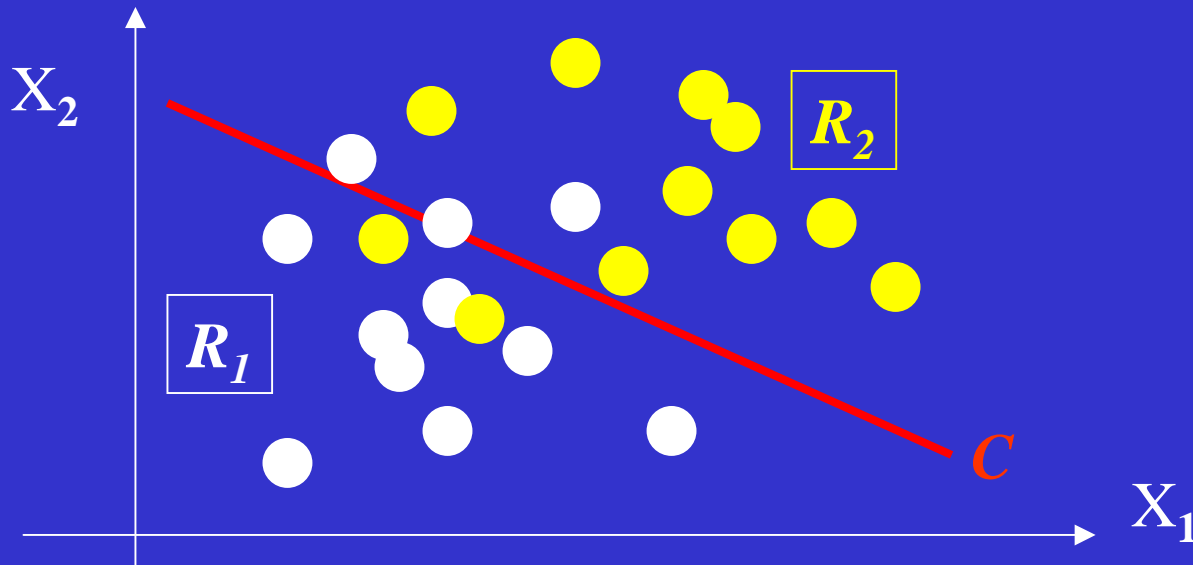
Basic Concept



Basic Concept(cont.)

Remark A :

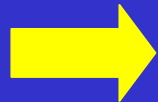
1. Classification rules cannot usually provide an “**error-free**” method of assignment, because the groups may **overlap**.
2. Possible to incorrectly classify a π_2 object as belonging to π_1 , and vice versa.



Basic Concept(cont.)

Remark B:

1. A **good** classification procedure should result in **few** misclassifications, *i.e.* the chances, or probabilities, of misclassification should be **small**.



Optimal classification rule

2. An **optimal** classification rule should take **prior probabilities of occurrence** and **cost** into account.

Basic Concept(cont.)

Remark B:

3. It may be that one class or population has a greater likelihood of occurrence than another because one of the two populations is relatively much larger than the other.

Prior distribution

4. Failing to diagnose a potentially fatal illness is substantially more costly than concluding the disease is present when, in fact, it is not.

Cost

Examples

1. Hemophilia A data(Taiwan)

Normal woman v.s. Definite carrier

by factor VIII coagulant activity & related antigen

2. Hemophilia A data

Noncarrier v.s. Obligatory carrier

by AHF activity & AHF antigen

3. Salmon data

Alaskan v.s. Canadian by Gender, Freshwater & Marine

Examples(cont.)

4. Iris data(Fisher)

Setosa v.s. Versicolor v.s. Virginica

by Sepal length, Sepal width, Petal length
& Petal width

5. Human acute leukemias data

Acute myeloid leukemia(AML) v.s. Acute lymphoblastic leukemia(ALL) by gene expression

6. Small round blue cell tumors(SRBCT) of childhood data

Burkitt lymphoma(BL) v.s. Ewing sarcoma(EWS) v.s.

Neuroblastoma(NB) v.s. Rhabdomyosarcoma(RMS)

by gene expression

Discrimination Models

- Linear discriminant analysis (LDA)

$$\text{Model : } l(X) = i_0 + i_1 X$$

Assumption : all the groups have equal covariance matrices
(homoscedastic model)

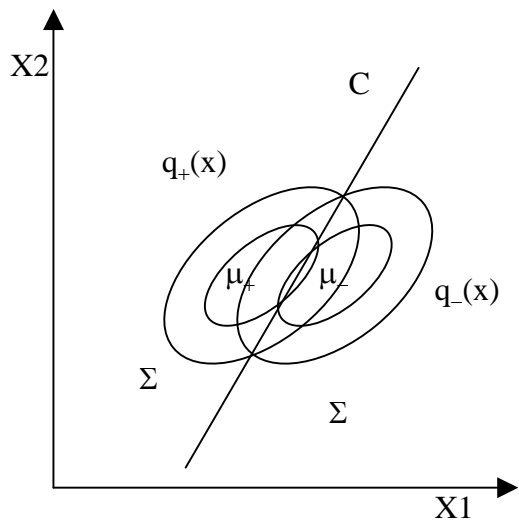
- Quadratic discriminant analysis (QDA)

$$\text{Model : } q(X) = i_0 + i_1 X + X' i_2 X$$

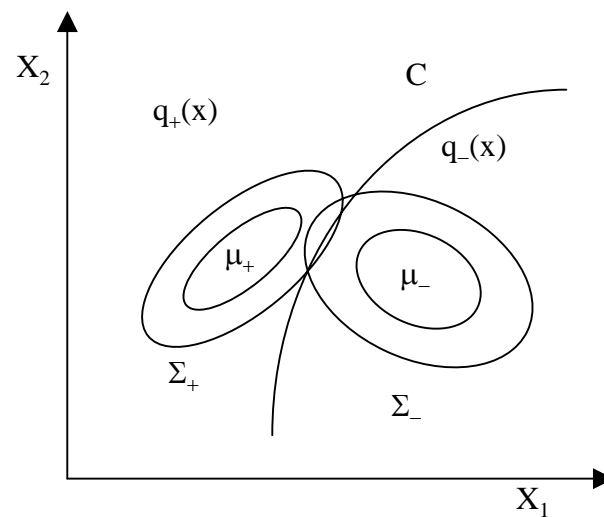
Assumption : the various groups have independent
covariance matrices
(heteroscedastic model)

LDA & QDA

Linear decision boundary for normal distributions when $\Sigma_+ = \Sigma_-$ (LDA)



Quadratic decision boundary for normal distributions when $\Sigma_+ \neq \Sigma_-$ (QDA)



Recent Methods---Classification

- Statistical methods(Fisher):
 - Golub, T., et al.(1999) Science
 - Hedenfalk, I., et al.(2001) N. Engl. J. Med.
 - Hastie, T., et al.(2001) Genome Biol.
- Artificial Neural networks:
 - Khan, J., et al.(2001) Nat. Med.
- Classification and Regression Tree(CART), Logistic
- Support Vector Machines(SVM)
- Boosting, Bagging
- Bayseian Logistic Regression with SVD

Challenge

- There are a large number of genes from which to predict classes (and they tend to be highly correlated) and a relatively small number of samples
- It is important to identify which genes contribute most to the classification
- **Discrimination** refer to defining previously unrecognized cancer type
- **Classification** refer to the assignment of particular cancer samples to already-defined classes, which could reflect current states or future outcomes

Summary

What is an appropriate method?

Thank You !