Clustering Algorithm

- **k-means**: $k$ clusters; each cluster is represented by the center of the cluster
- **PAM**: $k$ clusters; each cluster is represented by one of the objects in the cluster
- **Hierarchical clustering**: returns a complete tree with individual patterns as leaves and the convergence points of all branches as the root.
- **SOM**
SOM: Motivation

• Misleading dendrograms:

• The SOM clustering is designed to create a plot in which similar patterns are plotted next to each other.
Self-Organizing Feature Maps (SOM)

- SOM: A *map* consists of many simple elements (nodes or neurons); it is constructed by training.
  - SOMs are believed to resemble processing that can occur in the brain
  - Useful for visualizing high-dimensional data in 2- or 3-D space
  - The number of groups = number of nodes
1. Six nodes \((N_1, N_2, \ldots, N_6)\) of \(3 \times 2\) grids on 2D are first selected.

2. At each iteration, a data point (gene) \(P\) is randomly selected. \(N_p\) is the node that maps nearest to \(P\). Then the mapping of nodes is updated by moving points (all nodes) towards \(P\). \(N_p\) is moved the most.

3. Data point \(P\) is recycled and the procedures continue for 20,000-50,000 iterations.
Self-Organizing Feature Maps (SOM)

- This process can be visualized by imagining all SOM units being connected to each other by rubber bands.

A 2D SOFM trained on 3-dimensional data.
Example - Tamayo et al. (1999)

6 x 5 SOM.
The 828 genes that passed the variation filter were grouped into 30 clusters.

Each cluster is represented by the centroid (average pattern) for genes in the cluster.

Expression levels are shown on y-axis and time points on x-axis. Error bars indicate the SD of average expression. n indicates the number of genes within each cluster.
• Literature:
  – Algorithmic Approaches to Clustering Gene Expression Data
    [http://citeseer.nj.nec.com/shamir01algorithmic.html](http://citeseer.nj.nec.com/shamir01algorithmic.html)
R: Clustering Algorithm

• Partitioning methods (PM):
  – k-means: kmeans(stats)
  – PAM: pam(cluster)

• Hierarchical clustering (HC):
  – hclust(stats), agnes(cluster), diana(cluster)

• SOM
  – som(som)

• Visualization:
  – Silhouette plot: silhouette(cluster)
  – Reordering heatmap for HC: heatmap(stats), heatmap.2(gplots)
  – (R-2.7.0) heatmaps for PM and SOM: heatmapsM(maigesPack)
Example: Apop.xls

http://homepage.ntu.edu.tw/~lyliu/IntroBioinfo/Apop.xls

save the file as comma delimited (.csv).

> Apop = as.matrix(read.csv("Apop.csv",row.names=1))
> Apop = t(as.matrix(read.csv("Apop.csv",row.names=1)))
Partitioning Methods

kmeans:
> out.km = kmeans(Apop,3)

Available components:
[1] "cluster"  "centers"  "withinss"  "size"

PAM:
> library(cluster)
> out.pam3 = pam(Apop,3)

Available components:
[1] "medoids"  "id.med"  "clustering"  "objective"  "isolation"
[6] "clusinfo"  "silinfo"  "diss"  "call"  "data"
> si.pam3 = silhouette(out.pam3)
> plot(si.pam3)
Average Silhouette

• For each gene j, compute its silhouette ($S_j$):

$$S_j = \frac{b_j - a_j}{\max(a_j, b_j)}$$

- $a_j =$ average distance between gene j and other elements in the same group
- $b_j = \max_k b_{jk}$
- $b_{jk} =$ average distance between gene j and the elements in the $k$th group ($k \neq j$)

• Average silhouette = $\frac{1}{n} \sum_{j=1}^{n} S_j$
Silhouette Plot

Each observation is represented by a horizontal bar

Silhouette plot of pam(x = ruspini, k = 4)

n = 75

4 clusters C_i
j | n_j | avg_{i\in C_j} s_i
1 | 20 | 0.73
2 | 23 | 0.75
3 | 17 | 0.67
4 | 15 | 0.80

Average silhouette width = 0.74
Average Silhouette

• Number of clusters, $k$:
  For different $k$, compute the average silhouette; the largest average silhouette gives the optimal number of clusters.
Silhouette Plots for Different $k$

par(mfrow=c(2,2))
for(i in 2:5) {
  plot(silhouette(pam(Apop, i)),
       main = paste("k = ", i), do.n.k=FALSE, cex.names=0.5)
}

將視窗分成 $2 \times 2 = 4$ 個小區
For the different values of $k$:

- **$k = 2$**
  - Average silhouette width: 0.27
  - $j : n_j | \text{ave}_{c \in C_j} s_i$
    - 1: 30 | 0.26
    - 2: 7 | 0.33

- **$k = 3$**
  - Average silhouette width: 0.15
  - $j : n_j | \text{ave}_{c \in C_j} s_i$
    - 1: 16 | 0.15
    - 2: 6 | 0.28
    - 3: 15 | 0.11

- **$k = 4$**
  - Average silhouette width: 0.12
  - $j : n_j | \text{ave}_{c \in C_j} s_i$
    - 1: 14 | 0.08
    - 2: 6 | 0.28
    - 3: 6 | 0.06
    - 4: 11 | 0.13

- **$k = 5$**
  - Average silhouette width: 0.14
  - $j : n_j | \text{ave}_{c \in C_j} s_i$
    - 1: 12 | 0.12
    - 2: 5 | 0.11
    - 3: 6 | 0.08
    - 4: 11 | 0.11
    - 5: 3 | 0.53
Save the plot (LaTeX user):

```r
postscript("silhouette_Apop.ps")
par(mfrow=c(2,2))
for(i in 2:5) {
    plot(silhouette(pam(Apop,i)),
         main = paste("k = ",i),
         do.n.k=FALSE,
         cex.names=0.5)
}
dev.off()
```

GSview: ftp://mirror.cs.wisc.edu/pub/mirrors/ghost/ghostgum/gsv48w32.exe
Save the plot (MS Office Word user):

```r
win.metafile("silhouette_Apop.emf")
par(mfrow=c(2,2))
for(i in 2:5) {
    plot(silhouette(pam(Apop,i)),
         main = paste("k = ",i), do.n.k=FALSE,
         cex.names=0.5)
}
dev.off()
```
library(maigesPack)
heatmapsM(Apop,groups=out.km$cluster)
heatmapsM(Apop,groups=out.pam3$clustering)
Hierarchical Clustering

- **Bottom-up (agglomerative):**
  
  \[
  \texttt{hclust}(d, \textit{method})
  \]
  
  - \textit{d}: distance matrix
  - \textit{method}: "single", "complete", "average", "centroid"

  \[
  \texttt{agnes}(x, \textit{metric, method})
  \]
  
  - \textit{x}: data matrix or distance matrix
  - \textit{metric}: "euclidean", "manhattan"
  - \textit{method}: "single", "complete", "average", "centroid"

- **Top-down (divisive):**

  \[
  \texttt{diana}(x, \textit{metric})
  \]
  
  - \textit{x}: data matrix or distance matrix
  - \textit{metric}: "euclidean", "manhattan"
  - \textit{method}: "single", "complete", "average", "centroid"
Bottom-up: $hclust$ (stats)

Default

hang = -1
Bottom-up: hclust (stats)
Bottom-up: hclust (stats)

Default color

Brewer color (RdBu)
Bottom-up: agnes (cluster)

Banner of agnes(x = Apop)

Dendrogram of agnes(x = Apop)

Height

Agglomerative Coefficient = 0.5
Bottom-up: agnes (cluster)
Top-down: diana (cluster)

Banner of diana(x = Apop)

Dendrogram of diana(x = Apop)

Divisive Coefficient = 0.64
SOM

\texttt{som(data, xdim, ydim)}

\texttt{> library(som)}
\texttt{> som(Apop, 2, 3)}
Note: The data contains 6601 genes, measured at 18 time points.
Detecting differentially expressed genes in microarray data
Introduction

• In many cases, the purpose of microarray experiment is to compare the gene expression levels in two or several predetermined classes.
  – The comparison is often performed under gene-by-gene basis.
  – However, the genes are rarely independent.
  – For the convenient interpretability, differentially expression analysis usually ignore the dependencies between genes.
Fold Change

• Fold change is the important and intuitive approach to find differentially regulated genes:

\[
\text{Fold change (FC)} = \frac{\text{Expression of Experimental Sample}}{\text{Expression of Reference Sample}}
\]

\[
\log_2(\text{FC}) = \log_2(\text{Expression of experimental sample}) - \log_2(\text{Expression of reference sample})
\]
Fold Change

- Histogram of $\log_2(\text{fold-change})$:

Selects genes in the tails of the histogram by setting thresholds at the desired minimum fold change. For example, $\text{FC} > 2^{0.5} \rightarrow \log_2(\text{FC}) > 0.5$
Fold Change

- Fold change method can also be visualized on scatter plots and MA-plots.
Fold Change

• It may be the only possibility in cases where no, or very few replicates, are available.

• The fold change is chosen arbitrarily and cannot access the level of significance.

⇒ statistical tests!
# Standard Statistical Tests

Decide which genes are significantly regulated in a microarray experiment.

<table>
<thead>
<tr>
<th>Microarray Data</th>
<th>Paired data</th>
<th>Unpaired data</th>
<th>Complex data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>z-test</em></td>
<td><em>two-sample t-test</em></td>
<td><em>One-Way Analysis of Variance (ANOVA)</em></td>
</tr>
<tr>
<td><strong>Parametric Hypothesis Testing</strong></td>
<td><strong>t-test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assumptions and Test for Normality</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>QQplot</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Shapiro-Wilk Normality Test</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Non-Parametric Hypothesis Testing</strong></td>
<td><em>Wilcoxon signed-rank test</em></td>
<td><em>Wilcoxon rank-sum test, (Mann-Whitney U test).</em></td>
<td><em>Kruskal-Wallis test</em></td>
</tr>
</tbody>
</table>
Terminology in Hypothesis Testing

• The null hypothesis:
  – $H_0: \mu = 1.15$. (the average price of a gallon of gas is $1.15$)

• The alternative hypothesis:
  – $H_1: \mu > 1.15$. (gas prices were actually higher)
  – $H_1: \mu < 1.15$.
  – $H_1: \mu \neq 1.15$. 

Terminology in Hypothesis Testing

• The significance level ($\alpha$) is related to the degree of certainty you require in order to reject the null hypothesis in favor of the alternative.
  
  – Decide in advance

  – Reject the null hypothesis if the probability of observing a more extreme result than your sampled one (p-value) is less than the significance level.

  – The probability of incorrectly rejecting the null hypothesis when it is actually true (Type I error) is $100(1-\alpha)$%.

  – If you need more protection from this error, then choose a lower value of $\alpha$. 
Terminology in Hypothesis Testing

• **P-value:**
  – Definition: \( P(\text{observing at least this level of differential gene expression by random chance}) \)
  – The smaller the p-value, the less likely it is that the observed data have occurred by chance, and the more significant the result.
Terminology in Hypothesis Testing

- **Confidence intervals**: a range of values that have a chosen probability of containing the true hypothesized quantity.
  - Suppose, in our example, 1.15 is inside a 95% confidence interval for the mean, $\mu$. That is equivalent to being unable to reject the null hypothesis at a significance level of 0.05.
  - Conversely if the $100(1-\alpha\%)$ confidence interval does not contain 1.15, then you reject the null hypothesis at the alpha level of significance.
Steps of Hypothesis Testing

1. Determine the null and alternative hypothesis, using mathematical expressions if applicable.

2. Select a significance level (\( \alpha \)).

3. Take a random sample from the population of interest.

4. Calculate a test statistic from the sample that provides information about the null hypothesis.

5. Decision
   - If the value of the statistic is consistent with the null hypothesis then do not reject H0.
   - If the value of the statistic is not consistent with the null hypothesis, then reject H0 and accept the alternative hypothesis.
Hypothesis Testing in Microarray Study

• In all of the Microarray datasets, we are interested in identifying differentially expressed genes.
• The method would then be applied to every gene (one gene at a time) on the microarray in order to identify those genes that are differentially expressed.
• If the null hypothesis were true, then the variability in the data does not represent the biological effect under study, but instead results from difference between individuals or measurement error.
• We then select differentially expressed genes not on the basis of their fold ratio, but on the basis of their p-value.
Hypothesis Testing in Microarray Study

• **Hypothesis test for two groups:**
  – Two sample means: t-test (paired or independent)

• **Hypothesis test for more than two groups:**
  – One-Way Analysis of Variance (ANOVA)
Paired Data

• **Paired data**: there are two measurements from each object. We are interested in the difference between the two measurements.

Example: Samples are taken from 20 breast cancer patients, **before** and **after** a 16 week course of doxorubicin chemotherapy, and analyzed using microarray. There are 9216 genes.

⇒ Has a gene been up-regulated or down-regulated in breast cancer following doxorubicin chemotherapy?
Paired Data

• For each object, calculate the difference between the two measurements:

$$D_i = X_{i1} - X_{i2}$$

• The $D_i$’s can be viewed as a new set of 
  independent sample and can be tested whether 
  the population mean of $D_i$’s is equal to 0!

$$H_0: \mu_D = 0 \quad H_a: \mu_D \neq 0$$
Paired Data

Note that \( \frac{\bar{D} - \mu_D}{\sqrt{S_D^2 / n}} \sim t(n-1) \)

Under \( H_0: \mu_D = 0 \), \( t_0 = \frac{\bar{D} - 0}{\sqrt{S_D^2 / n}} = \frac{\bar{D}}{\sqrt{S_D^2 / n}} \sim t(n-1) \)

Reject \( H_0 \) if \( |t_0| < t_{\alpha/2,n-1} \) or if p-value < \( \alpha \)
Paired Data

Example (cont.): Gene ACAT2

\[
\overline{D} = 0.346955, \quad S_D^2 = 0.2315987
\]

\[
t_0 = \frac{\overline{D}}{\sqrt{S_D^2 / n}} = 3.2242, \quad p-value = 0.004465
\]

Reject \(H_0\)!

Note: we can rank the genes based on their p-values.
R: Paired Data

• Test by R:

```
t.test(x, y, paired = TRUE,
       alternative = c("two.sided", "less", "greater"))
```

```r
> dd = read.delim("perou.tab")
> ACAT2 = as.numeric(dd[which(dd$Gene == "ACAT2"),-1])
> t.test(ACAT2[1:20], ACAT2[21:40], paired=T)

Paired t-test

data:  ACAT2[1:20] and ACAT2[21:40]
t = -3.2242, df = 19, p-value = 0.004465
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5721855 -0.1217245
sample estimates:
mean of the differences
 -0.346955
```
Unpair Data

- **Unpaired data**: two measurements are taken from two objects independently.

Example: Samples are taken from 37 patients suffering from B-cell acute lymphoblastic leukemia (BCR/ABL) and 42 normal samples (NEG) and analyzed using Affymetrix arrays. There are 12625 genes.

⇒ We wish to identify the genes that are up- or down-regulated in BCR/ABL relative to NEG. (i.e., to see if a gene is differentially expressed between the two groups.)
Unpair Data

(1) if $\sigma_1^2 = \sigma_2^2 = \sigma^2$,

Statistic: \[
\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_p^2 (1/n_1 + 1/n_2)}} \sim t(n_1 + n_2 - 2)
\]

where \[
S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}
\]

(2) if $\sigma_1^2 \neq \sigma_2^2 \Rightarrow$ Welch’s Approximation!

Statistics: \[
\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t(\nu), \quad \nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^2}{n_1(n_1 - 1)} + \frac{S_2^2}{n_2(n_2 - 1)}}
\]
Unpaired Data

• To test whether $\sigma_1^2 = \sigma_2^2$:
  
  Compute $F_0 = \frac{s_1^2}{s_2^2}$; we claim that $\sigma_1^2 \neq \sigma_2^2$ if
  
  $$F_0 > F_{\alpha/2, n_1-1, n_2-1} \quad \text{or} \quad F_0 < F_{1-\alpha/2, n_1-1, n_2-1}$$
R: Unpaired Data

• Test for equal variance:
  \[ \text{var.test}(x, y) \]

• if \( \sigma_{12} = \sigma_{22} \):
  \[ \text{t.test}(x, y, \text{var.equal} = \text{TRUE}, \text{alternative} = \text{c("two.sided", "less", "greater")}) \]

• if \( \sigma_{12} \neq \sigma_{22} \):
  \[ \text{t.test}(x, y, \text{var.equal} = \text{FALSE}, \text{alternative} = \text{c("two.sided", "less", "greater")}) \]
> var.test(exprs(eset)[1,]~cl)

F test to compare two variances

data: exprs(eset)[1, ] by cl
F = 0.5856, num df = 36, denom df = 41, p-value = 0.1052
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.3100400 1.1209797
sample estimates:
ratio of variances
 0.5855576

> t.test(exprs(eset)[1,]~cl,var.eq=T)

Two Sample t-test

data: exprs(eset)[1, ] by cl
t = 0.7365, df = 77, p-value = 0.4637
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.07320496  0.15914468
sample estimates:
mean in group BCR/ABL mean in group NEG
  7.538354     7.495384

Note: we can rank the genes based on their p-values.
Assumption of t-test

• Normality assumption:
  – For paired t-test, it is the distribution of the subtracted data that must be normal.
  – For unpaired t-test, the distribution of both data sets must be normal.

• To test normality:
  – Visualization: normal probability plot
  – Hypothesis test: Shapiro-Wilk Normality Test

• If the assumption is not held ⇒ nonparametric methods!
BCR/ABL

Normal Q-Q Plot

Sample Quantiles

Theoretical Quantiles

p-value = 0.2441

NEG

Normal Q-Q Plot

Sample Quantiles

Theoretical Quantiles

p-value = 0.2548
Non-parametric Statistics

• Two good reasons to use non-parametric statistic.
  – Microarray data is noisy:
    • There are many sources of variability in a microarray experiment and outliers are frequent.
    • The distribution of intensities of many genes may not be normal.
    • Non-parametric methods are robust to outliers and noisy data.
  – Microarray data analysis is high throughput:
    • When analyzing the many thousands of genes on a microarray, we would need to check the normality of every gene in order to ensure that t-test is appropriate.
    • Those genes with outliers or which were not normally distributed would then need a different analysis.
    • It makes more sense to apply a test that is distribution free and thus can be applied to all genes in a single pass.
Wilcoxon Signed-Rank Test (paired data)

• **Hypothesis:** median(D) = 0.

• **Statistic:**

\[
z = \frac{T - \left[ \frac{n(n + 1)}{4} \right]}{\sqrt[24]{n(n + 1)(2n + 1)}} \sim N(0, 1) \text{ under } H_0
\]

\[T = \min(T^+, T^-)\]

\[T^+ = \text{sum of the ranks for the “positive” values}\]

\[T^- = \text{sum of the ranks for the “negative” values}\]
R: Wilcoxon Signed-Rank Test

- Test by R:

  ```r
  wilcox.test(x, y, paired = TRUE, alternative = c("two.sided", "less", "greater"))
  ```

```r
> wilcox.test(ACAT2[1:20], ACAT2[21:40], paired=T)

Wilcoxon signed rank test
data:  ACAT2[1:20] and ACAT2[21:40]
V = 33, p-value = 0.005581
alternative hypothesis: true location shift is not equal to 0
```
Wilcoxon Rank-Sum Test (unpaired data)

• Compute the rank sums:
  – Rank the observations in the combined sample from the smallest (1) to the largest (n1+n2)
  – \(T_1\) = the rank sum for samples 1
  – \(T_2\) = the rank sum for samples 2

• Statistic:
  \[U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - T_1\]
  \[U_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - T_2\]

  – one-tailed test statistic: \(U = U_1\)
  – two-tailed test statistic: \(U = \min(U_1, U_2)\)

\[Z = \frac{U - (n_1 n_2 / 2)}{\sqrt{n_1 n_2 (n_1 + n_2 + 1) / 12}} \sim N(0,1) \text{ under } H_0\]
R: Wilcoxon Rank-Sum Test

• Test by R:

    wilcox.test(x, y,
                alternative = c("two.sided", "less", "greater"))

> # rank-sum test
> wilcox.test(exprs(eset)[1,] ~ cl)

    Wilcoxon rank sum test

    data:  exprs(eset)[1, ] by cl
    W = 856, p-value = 0.4427
    alternative hypothesis: true location shift is not equal to 0
One-Way Analysis of Variance (ANOVA)

• The cases you need ANOVA:
  – when you need to compare more than two groups (e.g., drug 1, drug 2, and placebo)
  – when you need to compare groups created by more than one independent variable while controlling for the separate influence of each of them (e.g., Gender, type of Drug, and size of Dose).

• In fact, for two group comparisons, ANOVA will give results identical to a t-test.
One-Way Analysis of Variance (ANOVA)

- Example: ALL dataset

<table>
<thead>
<tr>
<th>Type</th>
<th>ALL1/AF4</th>
<th>BCR/ABL</th>
<th>E2A/PBX1</th>
<th>NEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>10</td>
<td>37</td>
<td>5</td>
<td>42</td>
</tr>
</tbody>
</table>

- We want to identify genes that are differentially expressed in one or more of these four groups.
### One-Way ANOVA

Let $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, 

- $i = 1, \cdots, p$  
- $j = 1, \cdots, n_i$.  

$\epsilon_{ij} \sim N(0, \sigma^2)$

$$\mu_j = \mu + \alpha_j$$  

$H_0: \mu_1 = \mu_2 = \cdots = \mu_p$  

Reject $H_0$ if $F_0 > F_{(\alpha, p-1, n-p)}$

---

The ANOVA Table for Comparing Means

<table>
<thead>
<tr>
<th>Source</th>
<th>SS (Sum of Squares)</th>
<th>DF</th>
<th>MS (Mean Square)</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>$SST = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\bar{y}_{ij} - \bar{y})^2$</td>
<td>$p-1$</td>
<td>$MST = \frac{SST}{p-1}$</td>
<td>$F_0 = \frac{MST}{MSE}$</td>
<td>$p$-value</td>
</tr>
<tr>
<td>Error</td>
<td>$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$</td>
<td>$n-p$</td>
<td>$MSE = \frac{SSE}{n-p}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$TSS = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$</td>
<td>$n-1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
> y = drop(exprs(eset[1,]))
> out = lm(y~factor(cl))
> anova(out)

Analysis of Variance Table

Response: y

         Df Sum Sq Mean Sq F value Pr(>F)
factor(cl) 3 0.2048  0.0683  1.0664  0.3675
Residuals 90 5.7604  0.0640
Assumption of ANOVA

- Two assumptions for the residuals (observed value – fitted value):
  - Normality assumption:
    - Visualization: normal probability plot
    - Hypothesis test: Shapiro-Wilk Normality Test
  - Equal variance:
    - Visualization: plot of residuals versus fitted values (means)
    - Hypothesis test: Bartlett’s Test

- If the assumption is not held ⇒ nonparametric methods!
Check Assumptions

plot(out, which=c(1:2))
Check Assumptions

```r
> shapiro.test(out$residuals)

       Shapiro-Wilk normality test

  data:  out$residuals
  W = 0.9759, p-value = 0.07968

> bartlett.test(out$residuals~cl)

     Bartlett test of homogeneity of variances

  data:  out$residuals by cl
Bartlett's K-squared = 3.2183, df = 3, p-value = 0.3592
```
Nonparametric ANOVA

Kruskal-Wallis Test:

> kruskal.test(y ~ factor(trt))

> kruskal.test(y ~ factor(cl))

Kruskal-Wallis rank sum test

data:  y by factor(cl)
Kruskal-Wallis chi-squared = 3.7234, df = 3, p-value = 0.2929
Comments

• The main hazard in using standard statistical tests occurs when there are too few replicates to obtain an accurate estimate of experimental variances. In such cases, modeling methods that use pooled variance estimates may be helpful.

• Standard interpretations of t and F tests assume that the data are sampled from normal populations with equal variances. Expression data may fail to satisfy either or both of these constraints.
Permutation Tests

- Permutation tests carried out by repeatedly scrambling the samples’ class labels and computing statistic for all genes in the scrambled data.

- Find the likelihood of the observed statistic based on the distribution of statistics from the permuted samples.
Permutation Tests

true class labels:

(random) permutations of class labels:

![Histogram of test statistics](image)

null distribution of test statistic
Permutation Tests

• **Step 1:** Permute the sample columns. Recalculate the statistic for the permuted sample.

• **Step 2:** Repeat Step 1 for all possible permutations.
  – # of permutations: \( B = \frac{n!}{n_1! \ n_2!} \)

• **Step 3:** Use the all permuted statistics to get the distribution

• **Step 4:** Get the p-value:
  – P-value = (# of permuted statistics the same as or more extreme than observed one) / B.
Permutation Tests

• Example:

<table>
<thead>
<tr>
<th>Class I</th>
<th>Class II</th>
<th>t-Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>e</td>
<td>2.1004</td>
</tr>
<tr>
<td>b</td>
<td>f</td>
<td>0.8431</td>
</tr>
<tr>
<td>c</td>
<td>g</td>
<td>-2.1004</td>
</tr>
<tr>
<td>d</td>
<td>h</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>a</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>7</td>
<td>2</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>a</th>
<th>f</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>b</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>2</td>
<td>15</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>h</th>
<th>d</th>
<th>f</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>10</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>a</th>
<th>c</th>
<th>g</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>15</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

#: # of possible combinations = 70
R: Permutation Tests

• “multtest” package:
  – mt.sample.teststat: to compute permuted statistics
    
    ```r
    > args(mt.sample.teststat)
    function (V, classlabel, test = "t", fixed.seed.sampling = "y",
             B = 10000, na = .mt.naNUM, nonpara = "n")
    ```

  – mt.sample.rawp: to compute the p-values
    
    ```r
    > args(mt.sample.rawp)
    function (V, classlabel, test = "t", side = "abs", fixed.seed.sampling = "y",
             B = 10000, na = .mt.naNUM, nonpara = "n")
    ```

Note: “test” includes

  t, t.equalvar, pairt, wilcoxon, f
Comment

• Generally best capture the unknown structure of the data.

• It is ideal when the number of arrays is sufficient to offer the desired degree of confidence.

• May be computationally expensive.
Bootstrap

• The bootstrap method attempts to determine the probability distribution from the data itself.

  **Step 1:** One computes a statistic from the current list.

  **Step 2:** Create an artificial list by randomly drawing elements from the current list. Some elements will be picked more than once.

  **Step 3:** Compute a new statistic.

  **Step 4:** Repeat 100-1000 times and look at the distribution of these objects.
Bootstrap

• Example (Hjorth, 1994):
  Eleven life lengths of an engine part were measured as
  
  \[
  5700 \quad 36300 \quad 12400 \quad 28000 \quad 19300 \quad 21500 \\
  12900 \quad 4100 \quad 91400 \quad 7600 \quad 1600
  \]

  **Step 1:** Estimate the population median by the sample median 
  \[
  \hat{\theta} = x_{(6)} = 12900
  \]
Bootstrap

Steps 2 & 3: Bootstrap simulations:

Table 5.1 Data drawn in the first five bootstrap samples.

<table>
<thead>
<tr>
<th>Original data ordered</th>
<th>Bootstrap sample number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1600</td>
<td>+</td>
</tr>
<tr>
<td>4100</td>
<td>++</td>
</tr>
<tr>
<td>5700</td>
<td>+</td>
</tr>
<tr>
<td>7600</td>
<td></td>
</tr>
<tr>
<td>12400</td>
<td>+</td>
</tr>
<tr>
<td>12900</td>
<td>+</td>
</tr>
<tr>
<td>19300</td>
<td>+</td>
</tr>
<tr>
<td>21500</td>
<td></td>
</tr>
<tr>
<td>28000</td>
<td>+</td>
</tr>
<tr>
<td>36300</td>
<td>+</td>
</tr>
<tr>
<td>91400</td>
<td>++</td>
</tr>
</tbody>
</table>

Table entries represent the presence of certain data points in different bootstrap samples. The table includes columns for bootstrap sample numbers 1 to 5. The bottom row shows the sample mean \( \hat{\theta}^* \) for each bootstrap sample, listed as 12900, 21500, 12900, 7600, and 12400.
Bootstrap

• After 200 simulations:
  average: 14843
  standard deviation: 5737
  bias = 14843 – 12900 = 1943
  A bias adjusted estimate of the population median: 12900 – 1942 = 10957

• This method can be applied to compute $p$-values:
  – P-value = (# of permuted statistics the same as or more extreme
    than observed one) / (Total # of simulations).
R: Bootstrap

```r
> library(boot)
> englife = c(5700, 36300, 12400, 28000, 19300,
+ 21500, 12900, 4100, 91400, 7600, 1600)
> boot.out = boot(englife, function(x, id) {median(x[id])}, 1000)
```

**ORDINARY NONPARAMETRIC BOOTSTRAP**

Call:
boot(data = englife, statistic = function(x, id) {
    median(x[id])
}, R = 1000)

Bootstrap Statistics :

    original  bias     std. error
    t1*   12900 2162.5 5861.693
How many bootstraps?

- No clear answer to this.
- Rule of thumb: try it 100 times, then 1000 times, and see if your answers have changed by much.
- Totally have $N^N$ possible subsamples.
Summary

- Non statistical method: fold change
- Standard statistical methods:
  - parametric
  - nonparametric
- Computation-intensive methods:
  permutation; bootstrap.
t-like tests:

Nonparametric rank-based statistics
References

Permutation tests:
- Dudoit, S., Yang, Y.-H., Callow, M.J. &