TransMedRi Workshop in Biostatistics: Critical evaluation of statistical analysis in scientific paper

Statistical methods in bioinformatical engineering

Ana Jerončić, PhD
University of Split - School of Medicine

What is BioInformatics?

- Sequence analysis and genome building
- Molecular structure prediction
- Evolution, phylogeny and linkage
- Automated data collection and analysis
- Simulations and modelling
- Biological databases and resources

Outline

1. Basic statistics
2. Hypothesis testing
3. Summary

Probability distributions

Progress

1. Basic statistics
2. Hypothesis testing
3. Summary

Statistical measures

Observations: $X_1, X_2, \ldots, X_n$

- Location measures, or
  Measures of central tendency
- Dispersion measures, or
  Measures of variability

Location

1. Mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$

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

3. Mode Mo

Example:

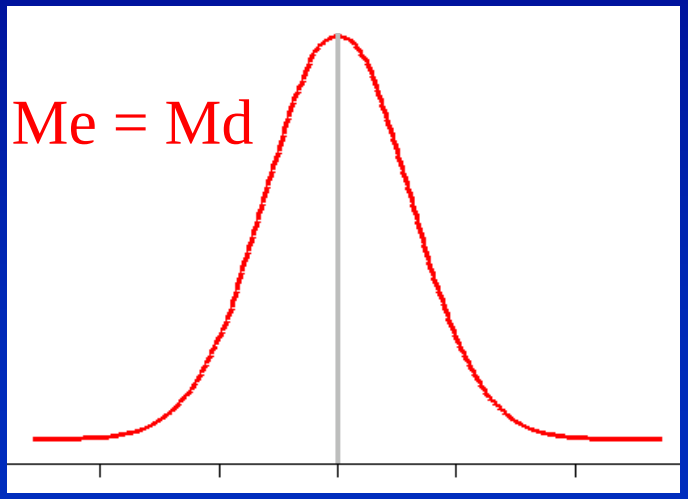
Observations : (1, 11, 10, 2, 7, 5)

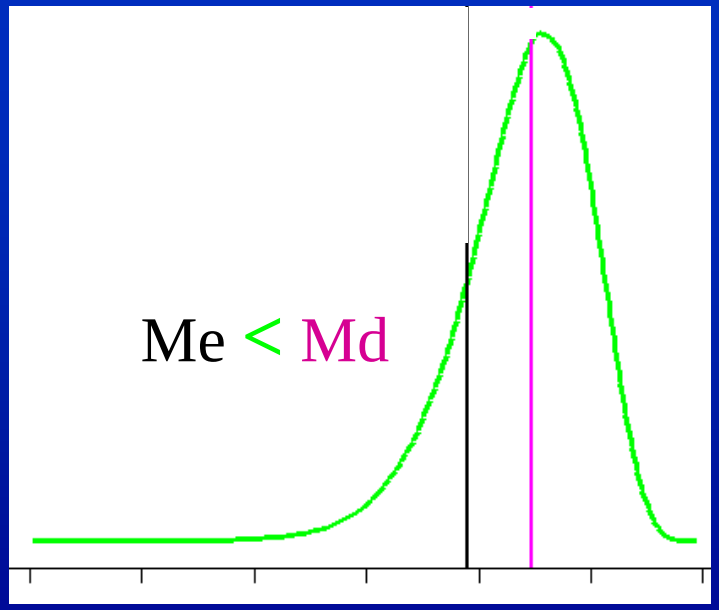
Order statistics : (1, 2, 5, 7, 10, 11)

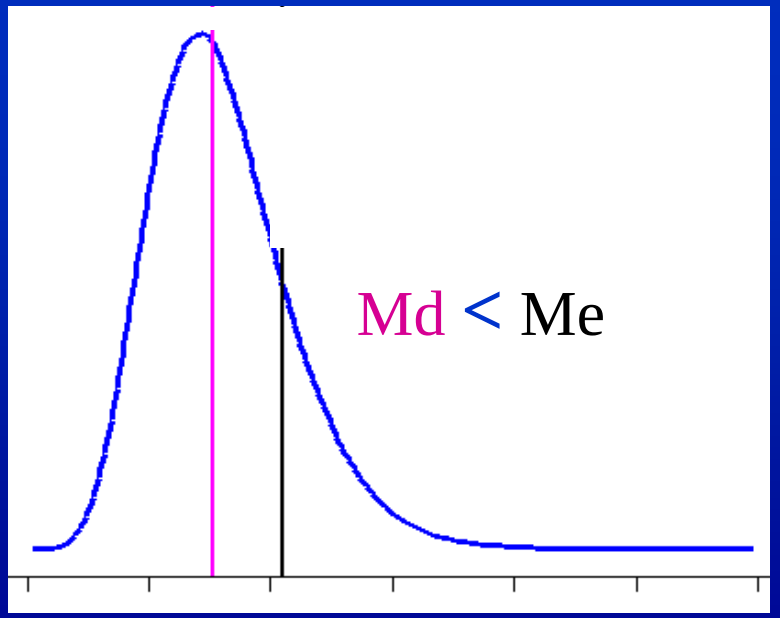
Mean : (1+11+10+2+7+5)/6 = 6

Median : $(X_{(3)}+X_{(4)})/2 = (5+7)/2 = 6$

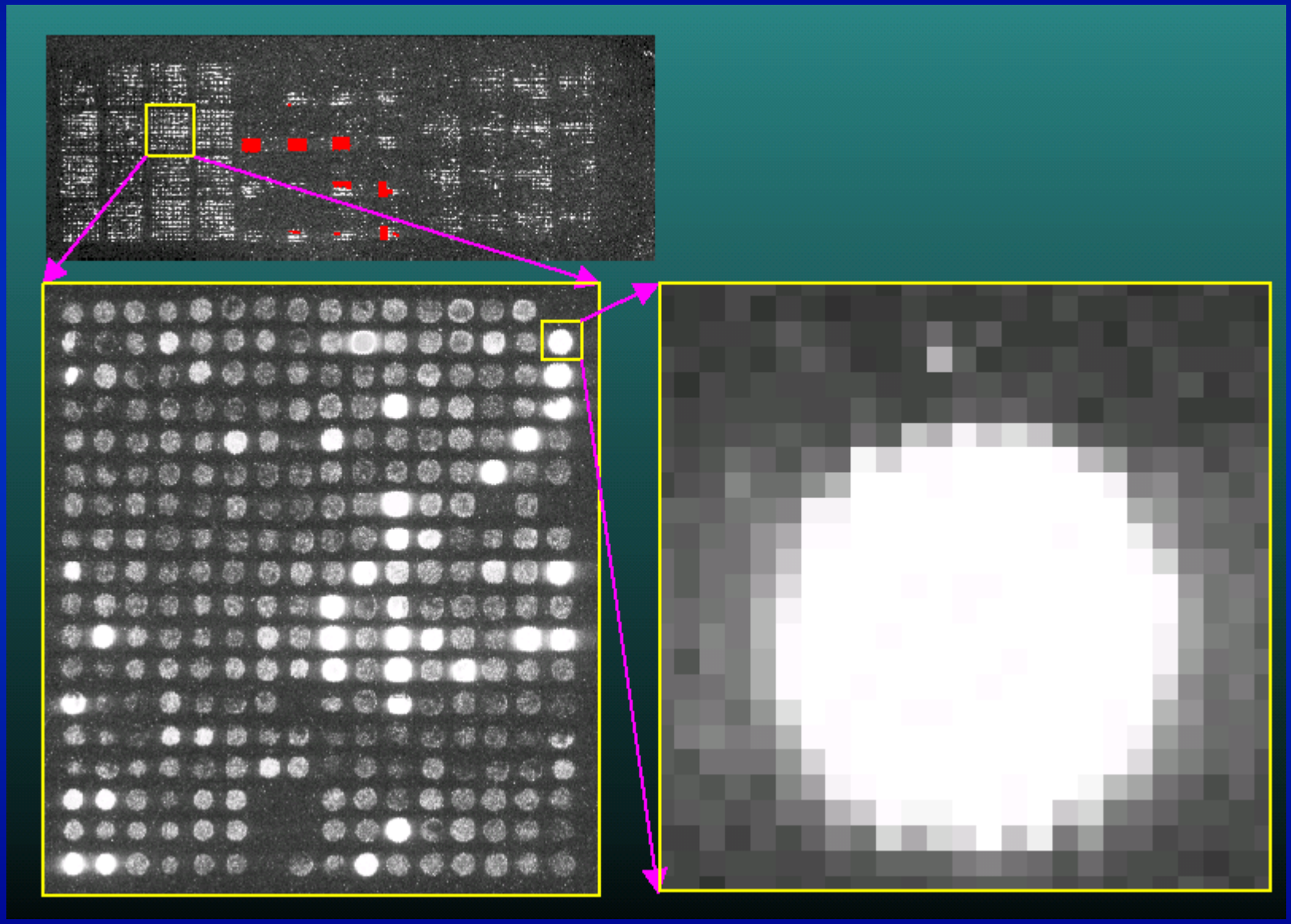
Mean v.s. Median





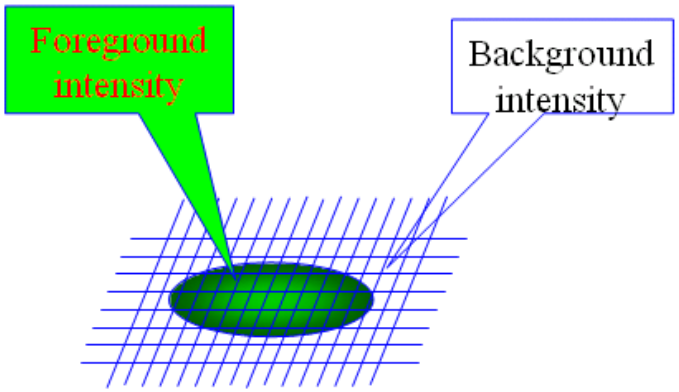


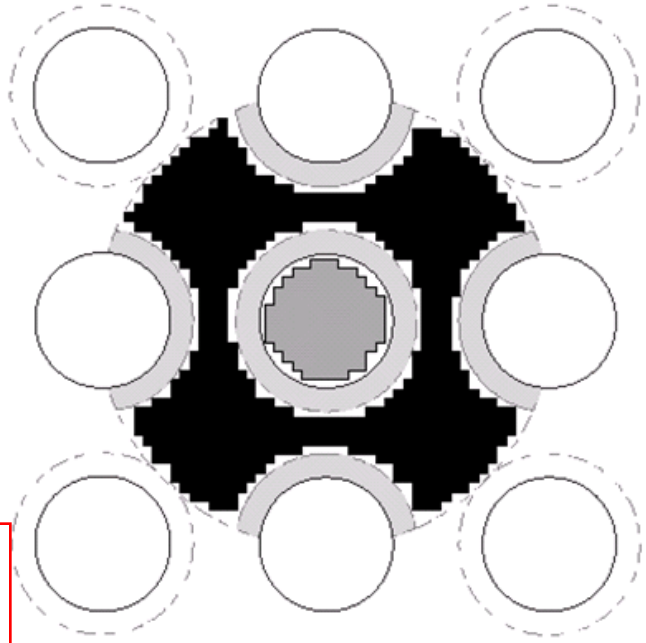
Example: Mean v.s. Median



Example: Mean v.s. Median

How to define background pixels and evaluate its intensity





Mean of pixels for foreground (*MeF*)
Median of pixels for foreground (*MdF*)

Mean of pixels for background (*MeB*)
Median of pixels for background (*MdB*)

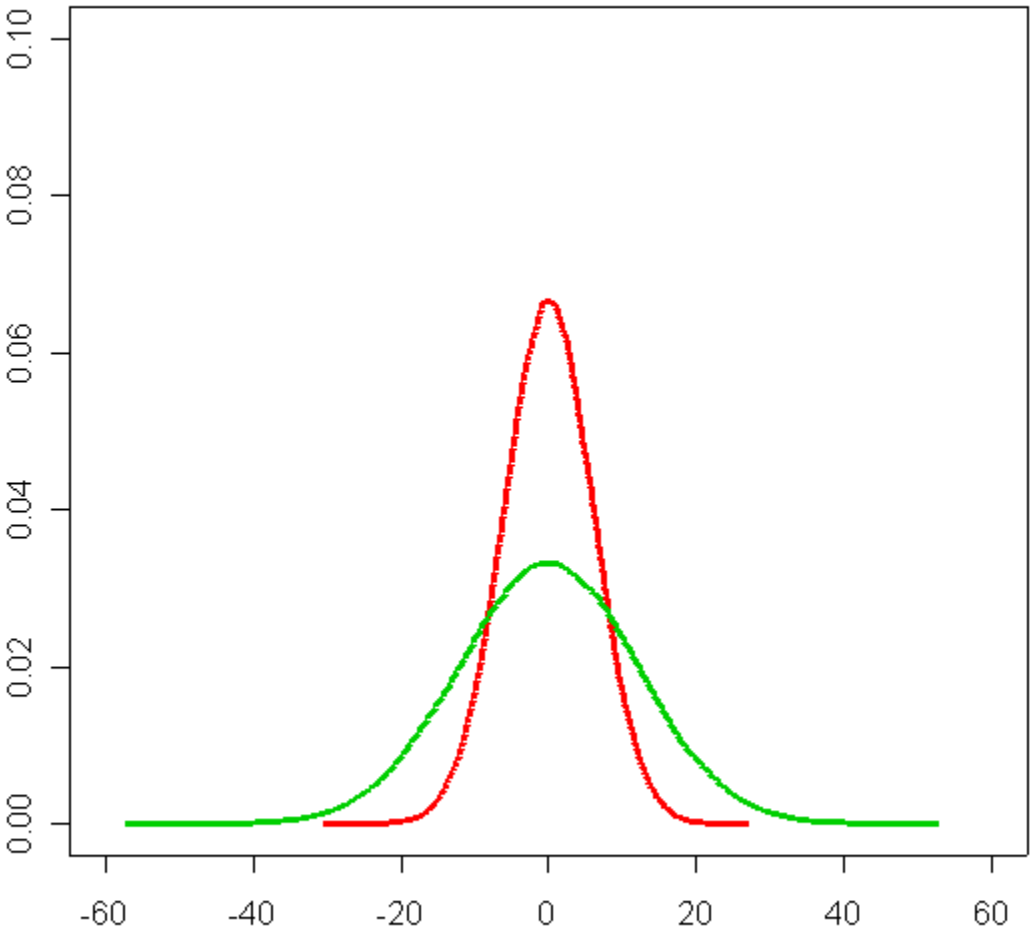
GenePix Pro5.0

Example: Mean v.s. Median

- The mean of pixels minus the mean of pixels (*MeF* - *MeB*)
- The median of pixels minus the median of pixels (*MdF* - *MdB*)
- The mean of pixels minus the median of pixels (*MeF* - *MdB*)
- The median of pixels minus the mean of pixels (*MdF* - *MeB*)

Dispersion

Dispersion



Dispersion

1. Range $R = X_{(n)} - X_{(1)}$
2. Interquartile-range
   $IQR = Q_3 - Q_1 = P_{75} - P_{25}$
3. Quartile deviation Q.D.=IQR/2

Example:

Observations : (1, 11, 10, 2, 7, 5)

Order statistics : (1, 2, 5, 7, 10, 11)

$R = X_{(6)} - X_{(1)} = 11 - 1 = 10$

$IQR = Q_3 - Q_1 = 10 - 2 = 8$

$Q.D. = IQR/2 = 8/2 = 4$

Dispersion

4. Mean Absolute Deviation

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right| \quad , \qquad MAD = \frac{1}{N}\sum_{i=1}^{N}|X_i - \mu|$$

5. Variance

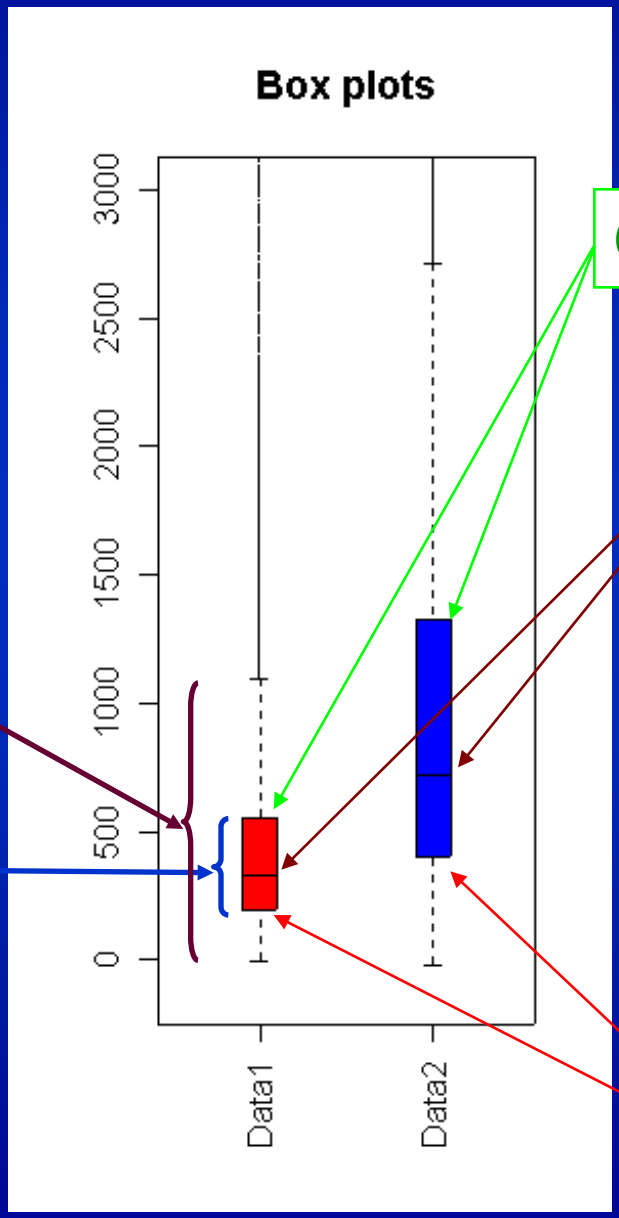
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \quad , \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^2$$

6. Standard Deviation

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right]}$$

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

Box plot



$Q_3=P_{75}$

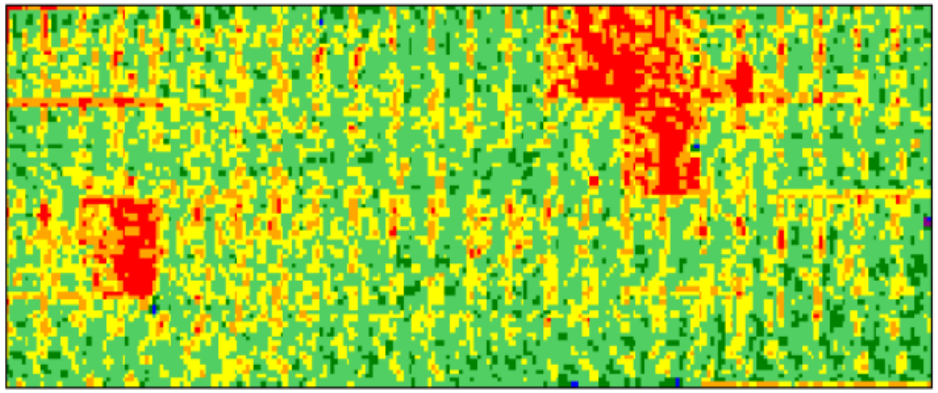
$Q_2=Md=P_{50}$

1.5IQR

$IQR=Q_3-Q_1$

$Q_1=P_{25}$

Detection of specific artifacts



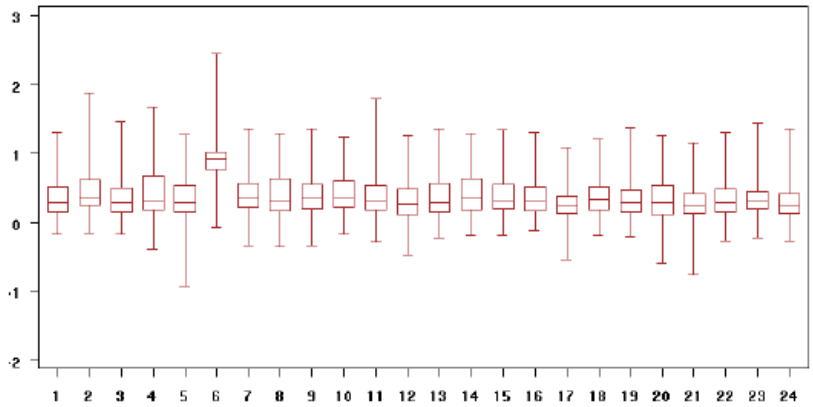
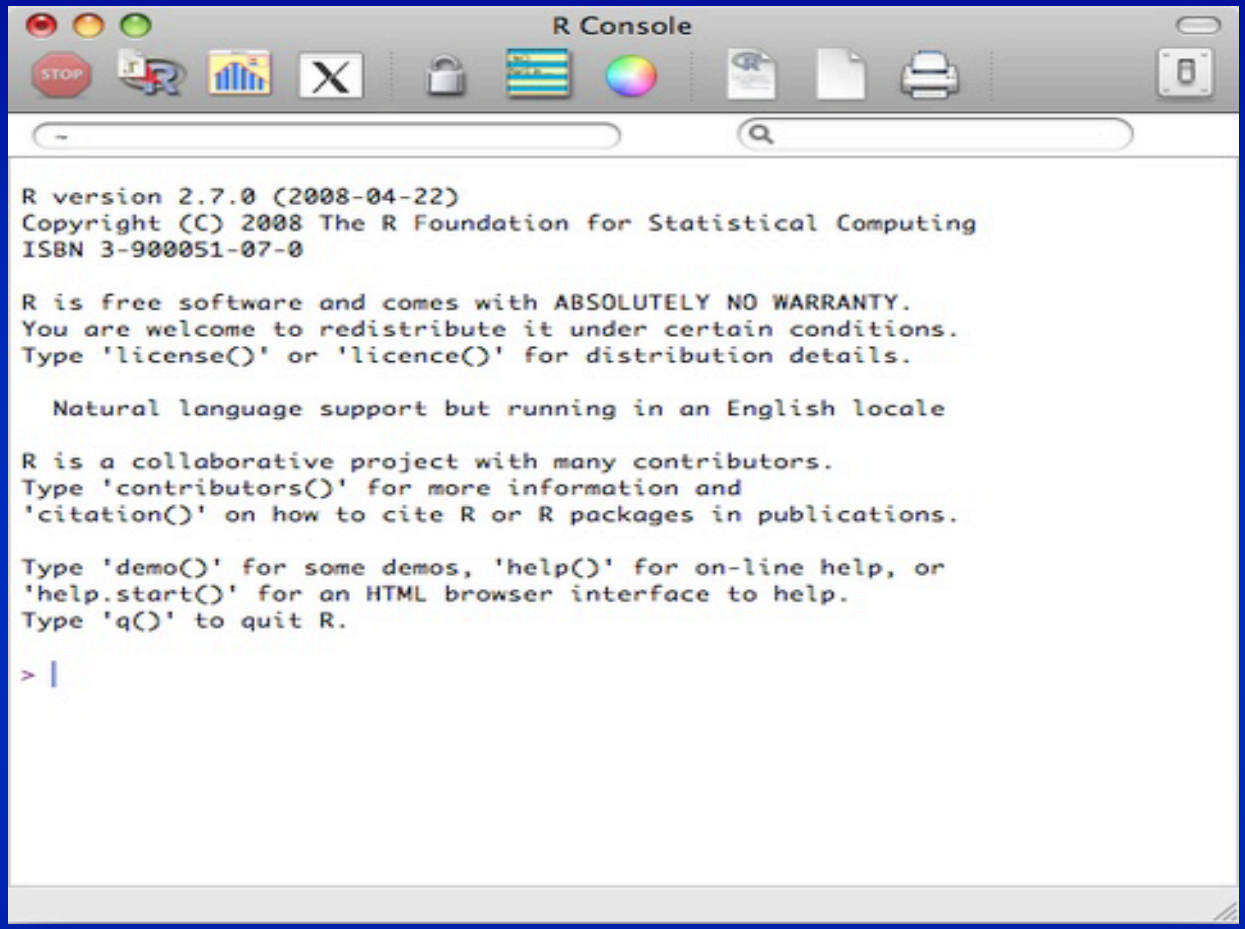
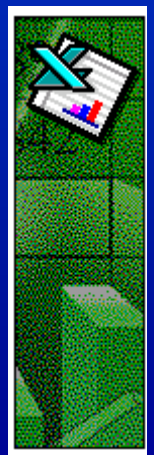


Figure 4: **Left:** Spatial distribution of the signal on the slide. Each pixel represents the uncorrected log-ratio of the median Cy5 (635 nm) and Cy3 (532 nm) channel fluorescence measurements, associated to a printed DNA feature. Background is not represented. Red squares correspond to print-tip effect. **Right:** Box plots per print-tip for the first 24 blocks of the previous slide. Print-tip 6 corresponds to the red square on the left of the slide.

Software for basic statistical analysis





sas

SPSS

R

S-PLUS

Discover S-PLUS — Revealing Graphics, Cutting-edge Statistics and Modeling

STATA Statistical Software for Professionals

Progress

1. Basic statistics
2. Hypothesis testing
3. Summary

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. $\mu \geq 2$ *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

Statistical 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.

1^0 $H_0 : \theta \geq \theta_0$ *v.s.* $H_a : \theta < \theta_0$

2^0 $H_0 : \theta \leq \theta_0$ *v.s.* $H_a : \theta > \theta_0$

3^0 $H_0 : \theta = \theta_0$ *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* $\overline{X}$

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

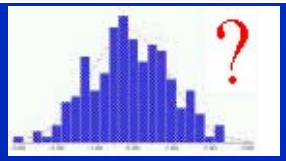
An Example

- A *hypothesis test* is a procedure for determining if an assertion about a characteristic of a population is reasonable.
- For example, suppose that someone says that the average price of a 1 L of gas in Croatia is 9.27 KN. How would you decide whether this statement is true?
  - You could try to find out what every gas station in the state was charging and how many litars they were selling at that price. That approach might be definitive, but it could end up costing more than the information is worth.
  - A simpler approach is to find out the price of gas at a small number of randomly chosen stations around the state and compare the average price to 9.27 KN.
- Of course, the average price you get will probably not be exactly 9.27 KN due to variability in price from one station to the next.
- Suppose your average price was 10.00 KN. Is this 73 lipas difference a result of chance variability, or is the original assertion incorrect?

A hypothesis test can provide an answer.

Hypothesis Testing

Decide which genes are significantly regulated in a microarray experiment.

Microarray Data	Paired data **Dependent samples**	Unpaired data **Independent samples**	Complex data *More than two Groups*
Parametric Hypothesis Testing	• z-test • ***t-test***	• ***two-sample t-test***	• One-Way Analysis of Variance (ANOVA)
	Assumptions and Test for Normality • Histogram, QQplot • Jarque-Bera test, Lilliefors test, Kolmogorov-Smirnov test		
Non-Parametric Hypothesis Testing	• Sign test, • Wilcoxon signed-rank test	• Wilcoxon rank-sum test, (Mann-Whitney U test).	

Breast Cancer Dataset (pair data)

- 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.
- **Paired data**: there are two measurements from each patient, one before treatment and one after treatment.
- These two measurements relate to one another, we are interested in the difference between the two measurements (the log ratio) to determine whether a gene has been up-regulated or down-regulated in breast cancer following doxoruicin chemotherapy.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D (2000)
Molecular portraits of human breast tumours. Nature 406:747-752.
Stanford Microarray Database:
http://genome-www.stanford.edu/breast_cancer/molecularportraits/

Leukemia Dataset (unpair data)

- Bone marrow samples are taken from 27 patients suffering from acute lymphoblastic leukemia (ALL) and 11 patients suffering from acute myeloid leukemia (AML) and analyzed using Affymetrix arrays. There are 7070 genes.
- **Unpaired data**: there are two groups of patients (ALL, AML).
- We wish to identify the genes that are up- or down-regulated in ALL relative to AML. (i.e., to see if a gene is differentially expressed between the two groups.)

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. et al. (1999)
Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531--537.
Cancer Genomics Program at Whitehead Institute for Genome Research
http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

Small Round Blue Cell Tumors (SRBCT) Dataset

- There are four types of small round blue cell tumors of childhood: neuroblastoma (NB), non-Hodgkin lymphoma (NHL), rhabdomyosarcoma (RMS) and Ewing tumors (EWS).
  - Sixty-three samples from these tumors, 12, 8, 20 and 23 in each of the groups, respectively, have been hybridised to microarray.
- We want to identify genes that are differentially expressed in one or more of these four groups.

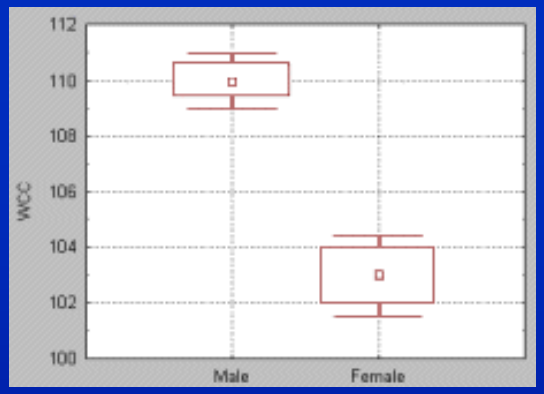
More on SRBCT:
http://www.thedoctorsdoctor.com/diseases/small_round_blue_cell_tumor.htm

Khan J, Wei J, Ringner M, Saal L, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu C, Peterson C and Meltzer P (2001)
Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7:673-679.
Stanford Microarray Database

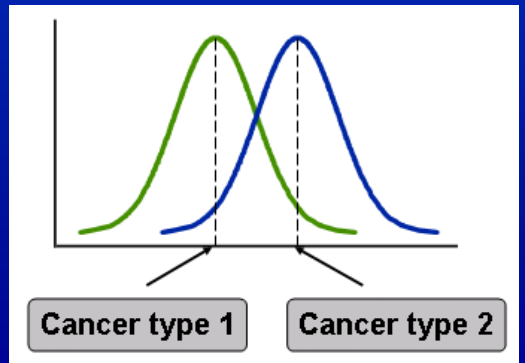
Terminology in Hypothesis Testing

- The ***null hypothesis:***
  - H0: μ = 9.27. (the average price of a 1L of gas is 9.27 KN)
- The ***alternative hypothesis:***
  - H1: μ > 9.27. (gas prices were actually higher)
  - H1: μ < 9.27.
  - H1: μ != 9.27.

$$\text{Power} = 1 - \beta.$$



Hypothesis Testing		Truth	
		H0	**H1**
Decision	**Reject H0**	Type I Error (alpha) (false positive)	Right Decision (true positive)
	Don't Reject H0	Right Decision	Type II Error (beta)



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 to reject the null hypothesis if the probability of observing your sampled result is less than the significance level.
  - For a typical significance level of 5%, the notation is alpha = 0.05. For this significance level, the chance of getting such or worse results under null hypothesis is 5%.
  - If you need more protection from this error, then choose a lower value of alpha .
- The ***p-value*** is the chance of observing the given or worse sample result under the assumption that the null hypothesis is true.
  - If the p-value is less than alpha, then you reject the null hypothesis.
  - For example, if alpha = 0.05 and the p-value is 0.03, then you reject the null hypothesis.

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, 9.27 is inside a 95% confidence interval for the mean, μ. 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 9.27, then you reject the null hypothesis at the alpha level of significance.

Rank Genes by p-Value

Definition:

p-Value is an estimate of the False Positive Rate (FPR)

= Prob(Decision=diff | Truth=not diff)

= Probability(|t-Statistic| ≥ |Observed t-Value|)

= Probability of Observing t-Values More Extreme

Property:

p-Value ≠False Discovery Rate

≈ False Positive Rate

= FP/(FP+TN)

Hypothesis Testing		Truth	
		H_0	H_1
Decision	Reject H_0	Type I Error (alpha) (false positive)	Right Decision (true positive)
	Don't Reject H_0	Right Decision	Type II Error (beta)

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.

Purpose of Hypothesis Testing

- 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.
- The smaller the p-value, the less likely it is that the observed data have occurred by chance, and the more significant the result.
- p=0.01 would mean there is a 1% chance of observing at least this level of differential gene expression by random chance.
- We then select differentially expressed genes not on the basis of their fold ratio, but on the basis of their p-value.

One Sample t-test

The One-Sample t-test compares the mean score of a sample to a known value. Usually, the known value is a population mean.
Assumption: the variable is normally distributed.

One sample t-test

$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$ (two-tailed).
μ: population mean.
α: significant level (e.g., 0.05).
Test Statistic:

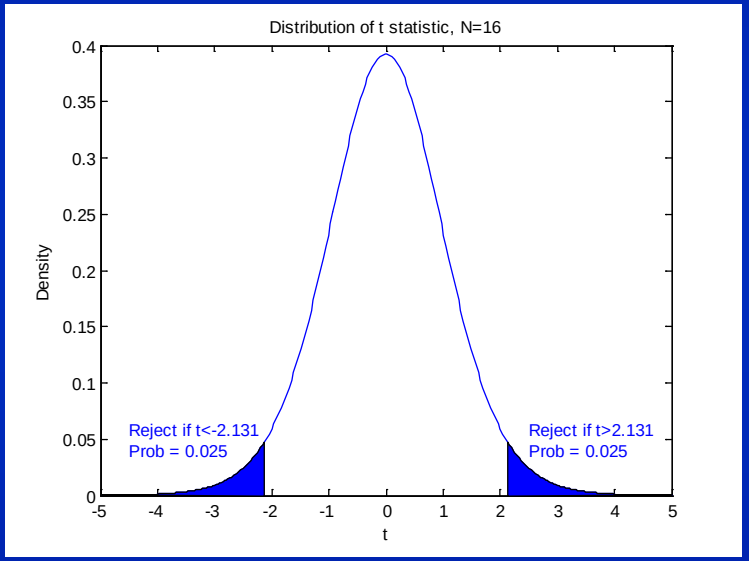
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \quad t_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$\bar{X}$: sample mean.

S: sample standard deviation.

n: number of observations in the sample.

- Reject H_0 if $|t_0| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ: $\bar{X} - t_{\alpha/2} S/\sqrt{n} \leq \mu < \bar{X} + t_{\alpha/2} S/\sqrt{n}$
- $p\text{-}value = P_{H_0}(|\mathbf{T}| > t_0)$, $\mathbf{T} \sim t_{n-1}$.



Two Sample t-test

Paired Sample t-test

$H_0 : \mu_d = \mu_0$
$H_1 : \mu_d \neq \mu_0$ (two-tailed).
μ_d: mean of population differences.
α: significant level (e.g., 0.05).
Test Statistic:

$$T_d = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}}, \quad t_d = \frac{\bar{d} - \mu_0}{S_d/\sqrt{n}}$$

$\bar{d}$: average of sample differences.

S_d: standard deviation of sample difference

n: number of pairs.

- Reject H_0 if $|t_d| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ_d: $\bar{d} - t_{\alpha/2}S/\sqrt{n} \leq \mu_d < \bar{d} + t_{\alpha/2}S/\sqrt{n}$
- $p\text{-}value = P_{H_0}(|\mathbf{T}| > t_d)$, $\mathbf{T} \sim t_{n-1}$.

Matlab syntax
h = ttest(x, m)
h = ttest(x, m, alpha)
[h, p, ci] = ttest(x, m, alpha, tail)

Two Sample t-test (Unpaired)

$H_0 : \mu_x - \mu_y = \mu_0$
$H_0 : \mu_x - \mu_y \neq \mu_0$
α: significant level (e.g., 0.05).
Test Statistic:

$$t_0 = \frac{(\bar{X} - \bar{Y}) - \mu_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

for homogeneous variances:
$df = n + m - 2$

for heterogeneous variances:
adjusted df

Reject H_0 if $|t_0| > t_{\alpha/2,\, df}$

Matlab syntax
h = ttest2(x, y)
[h, p, ci] = ttest2(x, y, alpha)
[h, p, ci, stats] = ttest2(x, y, alpha, tail, 'unequal')

Paired t-test Applied to A Gene From Breast Cancer Data

- The gene acetyl-Coenzyme A acetyltransferase 2 (ACAT2) is on the microarray used for the breast cancer data.
- We can use a paired t-test to determine whether or not the gene is differentially expressed following doxoruicin chemotherapy.
- The samples from before and after chemotherapy have been hybridized on separate arrays, with a reference sample in the other channel.
  - Perform the t-test. t=3.22 compare to t(19).
  - The p-value for a two-tailed one sample t-test is 0.0045, which is significant at a 1% significant level.

```
>> ACAT2 = importdata('ACAT2.txt');
>> ACAT2.diff = ACAT2.data(:,2)- ACAT2.data(:,3);
>> [h,pvalue,ci, stats] = ttest(ACAT2.diff)
```

```
h =

     1


pvalue =

    0.0045


ci =

    0.1215    0.5735


stats =

    tstat: 3.2186
       df: 19
       sd: 0.4828
```

- Conclude: this gene has been significantly down-regulated following chemotherapy at the 1% level.

Unpaired t-test Applied to A Gene From Leukemia Dataset

- The gene metallothionein IB is on the Affymetrix array used for the leukemia data.
  - To identify whether or not this gene is differentially expressed between the AML and ALL patients.
  - To identify genes which are up- or down-regulation in AML relative to ALL.
- t=-3.4177, p=0.0016

```
>> Leukemia = importdata('Leukemia.txt');
>> Met.IB.ALL = Leukemia.data((Leukemia.data(:,2)==1), 3);
>> Met.IB.AML = Leukemia.data((Leukemia.data(:,2)==0), 3);
>> [h,pvalue,ci,stats] = ttest2(Met.IB.ALL, Met.IB.AML, 0.01)
```

- Conclude that the expression of metallothionein IB is significantly higher in AML than in ALL at the 1% level.

```
h =

     1


pvalue =

    0.0016


ci =

   -1.8537   -0.2109


stats =

    tstat: -3.4177
       df: 36
       sd: 0.8444
```

Assumption of t-test

- The distribution of the data being tested is normal.
  - For paried 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.
- Homogeneous: the variances of the two population are equal. .
- Test for equality of the two variances: Variance ratio F-test.

Note:

- If the two populations are symmetric, and if the variances are equal, then the t-test may be used.
- If the two populations are symmetric, and the variances are not equal, then use the two-sample unequal variance t-test or Welch's t-test.

Non-parametric Statistics

- Do not assume that the data is normally distributed.
- There are 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.

Sign Test

- Given *n* pairs of data, the sign test tests the hypothesis that the median of the differences in the pairs is zero.
- The test statistic is the number of positive differences.
- If the null hypothesis is true, then the numbers of positive and negative differences should be approximately the same.
- In fact, the number of positive differences will have a Binomial distribution with parameters *n* and *p*.

Pair	Before	After	Sign
1	89	73	+
2	83	77	+
3	80	58	+
4	72	77	−
5	77	70	+
6	74	62	+
7	69	67	+
8	65	68	−
9	60	44	+
10	55	50	+
11	54	46	+
12	50	38	+
13	42	47	−
14	48	40	+
15	44	43	+
16	38	29	+
17	36	25	+

The Sign Test:

when $n_1 = n_2 \leq 50$

$H_0 : P = Q = \frac{1}{2}$

$H_1 : P \neq Q \neq \frac{1}{2}$

$T = \#"+"$

At $\alpha = 0.01$, two-tailed test,
reject H_0 if $T \geq 14$ when $N = 17$.
(Binomial Probability)

$\#"+" = 14$
$\#"-" = 3$

The obtained T=14 is equal to the critical value, so we reject H_0.

```
>> A=[89 83 80 72 77 74 69 65 60 55 54 50 42 48 44 38 36];
>> B=[73 77 58 77 70 62 67 68 44 50 46 38 47 40 43 29 25];
>> [p, h, stats]=signtest(A-B)
```

```
p =

    0.0127

h =

     1

stats =

    sign: 3
```

Wilcoxon Signed-Rank Test (paired)

- Null hypothesis: the population median from which both samples were drawn is the same.
- The sum of the ranks for the "positive" (up-regulated) values is calculated and compared against a precomputed table to a p-value.
  - Sorting the absolute values of the differences from smallest to largest.
  - Assigning ranks to the absolute values.
  - Find the sum of the ranks of the positive differences.
- If the null hypothesis is true, the sum of the ranks of the positive differences should be about the same as the sum of the ranks of the negative differences.

Pair	Before	After	Diff.	Rank
1	89	73	16	15.5
2	83	77	6	7
3	80	58	22	17
4	72	77	−5	5
5	77	70	7	8
6	74	62	12	13.5
7	69	67	2	2
8	65	68	−3	3
9	60	44	16	15.5
10	55	50	5	5
11	54	46	8	9.5
12	50	38	12	13.5
13	42	47	−5	5
14	48	40	8	9.5
15	44	43	1	1
16	38	29	9	11
17	36	25	11	12

The Wilcoxon signed-rank Test:

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$
$T = \min\{\sum_+ \text{Rank}, \sum_- \text{Rank}\}$
At $\alpha = 0.01$, two-tailed test,
reject H_0 if $T \leq 23$ when $N = 17$.
(Table)

(The zero difference is ignored when assigning ranks. $N_{new} = N_{old} - \#\{ties\}$)

$T = \min\{\sum_+ \text{Rank} = 140, \sum_- \text{Rank} = 13\}$
$= 13$

The obtained T=13 is less than the critical value 23, so we reject H_0.

```
>> A=[89 83 80 72 77 74 69 65 60 55 54 50 42 48 44 38 36];
>> B=[73 77 58 77 70 62 67 68 44 50 46 38 47 40 43 29 25];
>> [p, h, stats]=signrank(A, B)
```

```
p =

    0.0026


h =

     1


stats =

          zval: -3.0089
    signedrank: 13
```

One-Way Analysis of Variance (ANOVA)

- It often happens in research practice that you need to compare more than two groups (e.g., *drug 1*, *drug 2*, and *placebo*), or 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 these cases, you need to analyze the data using *Analysis of Variance*, which can be considered to be a generalization of the t-test.
- In fact, for two group comparisons, ANOVA will give results identical to a t-test.
- When the design is more complex, ANOVA offers numerous advantages that t-tests cannot provide (even if you run a series of t-tests comparing various cells of the design).
- Analysis of Variance (ANOVA) allows us to extend this to more than two populations or measurements (treatments). That is, we can test the following:
  - Are all the means from more than two populations equal?
  - Are all the means from more than two treatments on one population equal? (This is equivalent to asking whether the treatments have any overall effect.)

ANOVA Table

Treatment

	$i=1$	$i=2$	⋯	$i=p$	
	y_{11}	y_{21}		y_{p1}	
	y_{12}	y_{22}		y_{p2}	
	⋮	⋮	⋯	⋮	
	y_{1n_1}	y_{2n_2}		y_{pn_p}	Overall Mean
Means	$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$	⋯	$\bar{y}_{p\cdot}$	$\bar{y}$

One-Way ANOVA

$$y_{ij} = \mu + \alpha_{i\cdot} + \epsilon_{ij},$$

$$i = 1, \cdots, p.$$

$$j = 1, \cdots, n_i.$$

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

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

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

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

The ANOVA Table for Comparing Means

Source	SS *(Sum of Squares)*	DF	MS *(Mean Square)*	F	*Prob* > F
Treatment	$SST = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\bar{y}_{i\cdot} - \bar{y})^2$	*p-1*	$MST = \frac{SST}{p-1}$	$F_0 = \frac{MST}{MSE}$	*p*-value
Error	$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$	*n-p*	$MSE = \frac{SSE}{n-p}$		
Total	$TSS = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$	*n-1*			

ANOVA Applied to a Gene from (SRBCT) Dataset

```
>> ARP1 = importdata('ARP1.txt');
>> [p, table, stats] = anova1(ARP1.data(:,2), ARP1.data(:,1))
```

```
p =

  7.5096e-004


table =

    'Source'    'SS'          'df'    'MS'         'F'          'Prob>F'
    'Groups'    [ 2.7145]     [ 3]    [0.9048]     [6.4496]     [7.5096e-004]
    'Error'     [ 8.2772]     [59]    [0.1403]           []               []
    'Total'     [10.9917]     [62]          []           []               []


stats =

    gnames: {4x1 cell}
         n: [23 8 12 20]
    source: 'anova1'
     means: [1.1053 0.7280 1.4254 0.9617]
        df: 59
         s: 0.3746
```

Matlab syntax

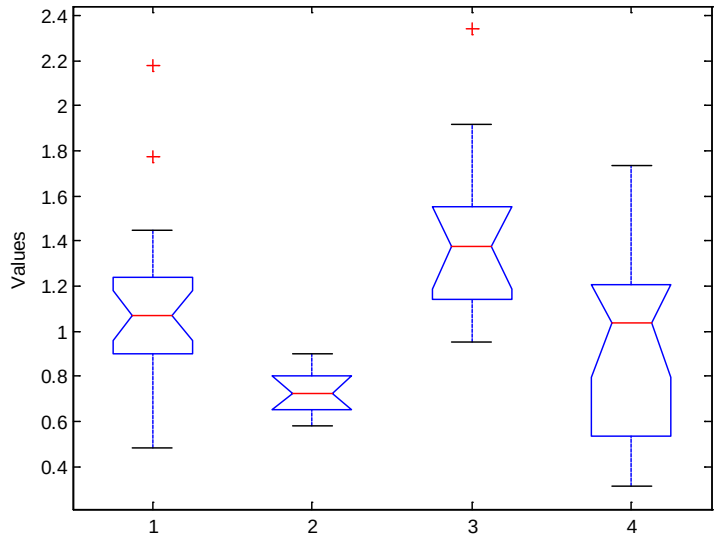
p = anova1(X)

p = anova1(X, group)

p = anova1(X, group, ‘displayopt’)

[p, table] = anova1(…)

[p, table, stats] = anova1(…)



Rank Genes by q-Value

Definition:

q-Value is an estimate of the False Discovery Rate (FDR)

= E(Truth=not diff | Decision=diff)

≈ p-Value × # of Genes / # p-Value Discoveries

(Benjamini, et. al., Bioinformatics, 2003:368-375)

Property:

q-Value ≠False Positive Rate

≈ False Discovery Rate

=Number of False Discoveries / Number of Discoveries

= FP/(FP+TP)

Hypothesis Testing		Truth	
		H_0	H_1
Decision	**Reject H_0**	Type I Error (alpha) (false positive)	Right Decision (true positive)
	Don't Reject H_0	Right Decision	Type II Error (beta)

Example (10000 Genes)

Declared Discoveries

Real Change	No	Yes	Row Totals
No	9025	475 False Positives	9500 475/9500 = .05
Yes	25	475	500
Column Totals	9050	950 475/950 = .50	10000

FPR ≈ p-Value

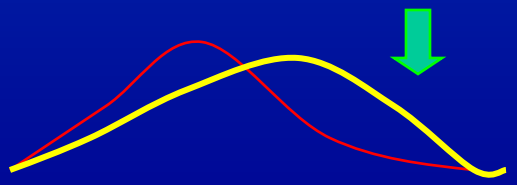
FDR ≈ q-Value

Multiple Testing Correction

- Family-wise Error Rate (FWER)
  - Control the probability of making ***any*** false positive call at the desired significance level. (FWER=prob(FP $\geq$ 1))
  - Conservative methods such as the Bonferroni correction
    - Divide p-value by number of calls (or genes)
    - For 12,000 genes, need a p-value threshold of about 0.000004 for each gene to assure that the probability of making any false present call is 0.05
- False Discovery Rate (FDR)
  - FDR = expected (# false predictions/ # total predictions)
  - Control the *proportion* of false positive calls in all positive calls at the desired significance level.
    - FDR chosen to maximize the number of genes passing the cutoff yet keeping the number of false calls small.
    - Shifts focus to predicted positives and accepts that some will be wrong. An FDR of 0.05 means out of 100 predicted positives, 5 are wrong.
    - FWER will be very high but not important.

FDR Using Permutations

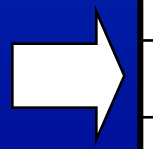
- The experiments are permuted between the groups and a permutation estimate of the null distribution of the statistic is compiled.
- This distribution is then used to estimate what cutoff value for the single gene *p*-values will achieve the desired FDR based on the number of genes expected to pass the cut-off by chance and the observed number of genes that pass the cut-off.



Using Permutations to Get p-values

- Permutations randomize associations
  - use these associations to recalculate statistics for each permutation
  - Find the likelihood of the observed statistic based on the distribution of statistics from the permuted samples
    - p-value = percent of statistics generated from the permutations that are greater than or equal to the observed statistic.
- Permuting columns preserves any gene dependencies but scrambles samples

LT-HSC				ST-HSC			
a	b	c	d	e	f	g	h
5	7	8	8	6	7	7	8
24	16	28	19	12	11	9	2
12	7	15	10	7	2	10	5



LT-HSC				ST-HSC			
c	d	e	f	a	b	g	h
8	8	6	7	5	7	7	8
28	19	12	11	24	16	9	2
15	10	7	2	12	7	10	5

Permute columns to c,d,e,f;a,b,g,h

Example of Permutation for Eight Samples in Two Groups (4, 4)

- Step 1: permute the sample columns (e.g., swap sample a in group 1 with the sample e in group 2). Calculate the t-statistic. *Note: permuting columns avoids assumption of gene independence. Important for considering more than one gene.*
- Step 2: repeat step 1 for all possible permutations (70 in this case).
  - B permutations are performed: B= n!/n1! n2!
- Step 3: Use the 70 t-statistics to get the distribution.
- Step 4: compare starting t-statistic to distribution to get p-value (fraction of permuted t-statistics larger than observed t-statistic *d*)

Generating Permutations

- 4, 4 (a,b,c,d; e,f,g,h) gives 70 permutations
  - No swap: (a,b,c,d; e,f,g,h)
    1
  - swap 1: (b,c,d,e; a,f,g,h) (a,c,d,e; b,f,g,h) (a,b,d,e; c,f,g,h) (a,b,c,e; d,f,g,h)
    16 (b,c,d,f; a,e,g,h) etc…
    (b,c,d,g; a,e,f,h) etc…
    (b,c,d,h; a,e,f,g) etc…
  - swap 2: (c,d,e,f; a,b,g,h) (b,d,e,f; a,c,g,h) (b,c,e,f; a,d,g,h) (a,d,e,f;b,c,g,h) (a,c,e,f; b,d,g,h) (a,b,e,f; c,d,g,h)
    36 (c,d,e,g; a,b,f,h) etc…
    (c,d,e,h; a,b,f,g) etc…
    (c,d,f,g; a,b,e,h) etc…
    (c,d,f,h; a,b,e,g) etc…
    (c,d,g,h; a,b,e,f) etc…
  - swap 3: (d,e,f,g; a,b,c,h) (c,e,f,g; a,b,d,h) (b,e,f,g; a,c,d,h) (a,e,f,g; b,c,d,h)
    16 (d,e,f,h; a,b,c,g) etc…
    (d,e,g,h; a,b,c,f) etc…
    (d,f,g,h; a,b,c,e) etc…
  - swap 4: (e,f,g,h; a,b,c,d)
    1

Available Hypothesis Tests in Matlab (statistics Toolbox)

1	Jarque-Bera test for goodness-of-fit to a normal distribution	`jbtest`
2	Lilliefors test for goodness of fit to a normal distribution	`lillietest`
3	Kolmogorov-Smirnov test of the distribution of one sample	`kstest`
4	Kolmogorov-Smirnov test to compare the distribution of two samples	`kstest2`
5	Hypothesis testing for the mean of one sample with known variance	`ztest`
6	Hypothesis testing for a single sample mean (paired)	`ttest`
7	Hypothesis testing for the difference in means of two samples (unpaired)	`ttest2`
8	Sign test for paired samples (paired)	`signtest`
9	Wilcoxon signed rank test of equality of medians (paired)	`signrank`
10	Wilcoxon rank sum test that two populations are identical (unpaired) (Mann-Whitney test)	`Ranksum`

11	One-Way Analysis of Variance (ANOVA)	`anova1`

Available Hypothesis Tests in R (Compared with Matlab)

R	Matlab
t.test(paired=T)	ttest
t.test(paired=F)	ttest2
wilcox.test(paired=T)	signrank
wilcox.test(paired=F)	Ranksum
anova	anova1

Progress

1. Basic statistics
2. Hypothesis testing
3. Summary

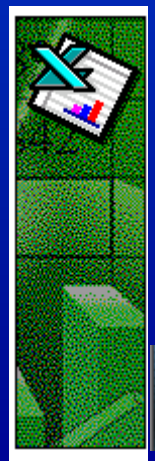
Summary

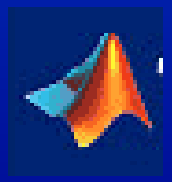
1. Inferential statistics

   t-test、Wilcoxon test、 Regression、F-test、Chi-square test、ANOVA、GEE...

2. Software

   Excel、SAS、SPSS、Stata、S-Plus、R、Matlab 、 ...




Thank you!