ANOVA and the Bootstrap

Ulrich Mansmann

mansmann@imbi.uni-heidelberg.de

Practical microarray analysis
March 2003
Heidelberg

Probe conservation and gene expression

Question of interest:

Does time between probe harvesting and probe hybridisation have a *relevant* influence on gene expression?
Are there degradation effects or time pattern?

Data:

Probes of three patients were hybridised at day 0, 1, and 2
Nine Affimetrix – chips with 12625 genes each

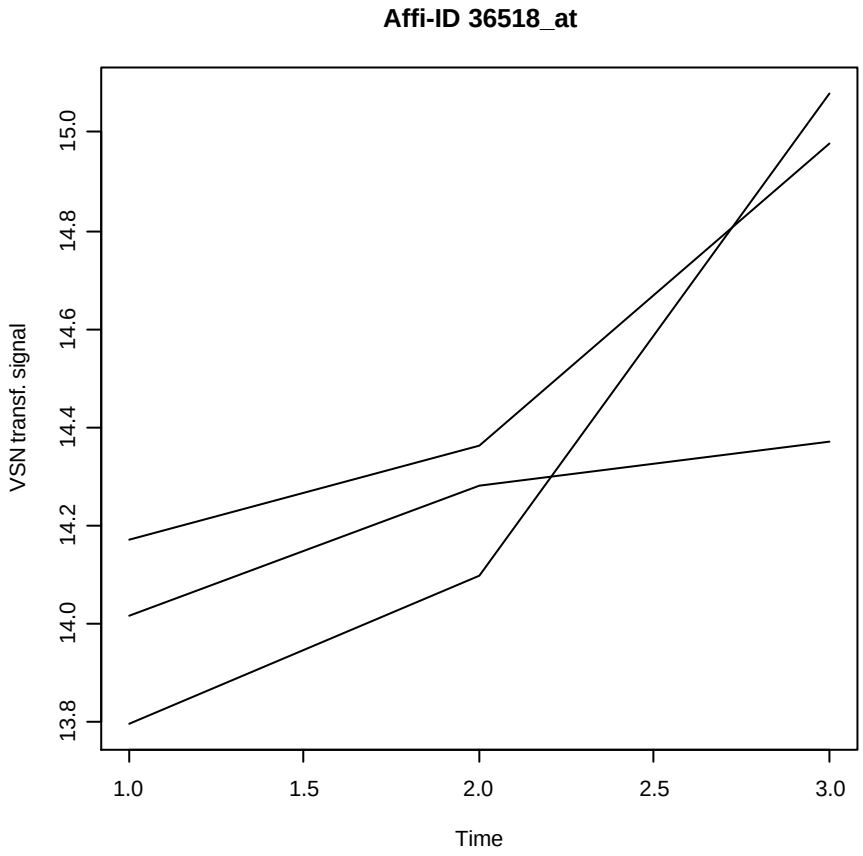
Methodological approach:

Quantify *time effects*: general or gene-specific
Is there evidence for time effects?
How to quantify variability between patients?
How to quantify variability over time (within patients)?
How to operationalise the idea of *relevant influence on gene expression)*

The one gene scenario - Table

time probe/patient	T_1	T_2	T_3
P_1	X_{11}	X_{12}	X_{13}
P_2	X_{21}	X_{22}	X_{23}
P_3	X_{31}	X_{32}	X_{33}

The one gene scenario - Graph



First ideas on variability

Quantifying variability: $SS = \sum_{i=1}^{3}\sum_{j=1}^{3}(x_{ij} - \bar{x})^2$ with $\bar{x} = \frac{1}{9}\sum_{i=1}^{3}\sum_{j=1}^{3} x_{ij}$, the global mean

Separate variability between patients from the variability of individual time courses (within patients)

$$SS = \sum_{i=1}^{3}\sum_{j=1}^{3}(x_{ij} - \bar{x})^2$$

$$= \sum_{i=1}^{3}\sum_{j=1}^{3}(x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 \text{ with } \bar{x}_i = \frac{1}{3}\sum_{j=1}^{3} x_{ij} \text{ mean patient value}$$

$$= \sum_{i=1}^{3}\sum_{j=1}^{3}(x_{ij} - \bar{x}_i)^2 + 3\sum_{i=1}^{3}(\bar{x}_i - \bar{x})^2$$

within patient variability between patient variability

ANOVA – Analysis of variance (1)

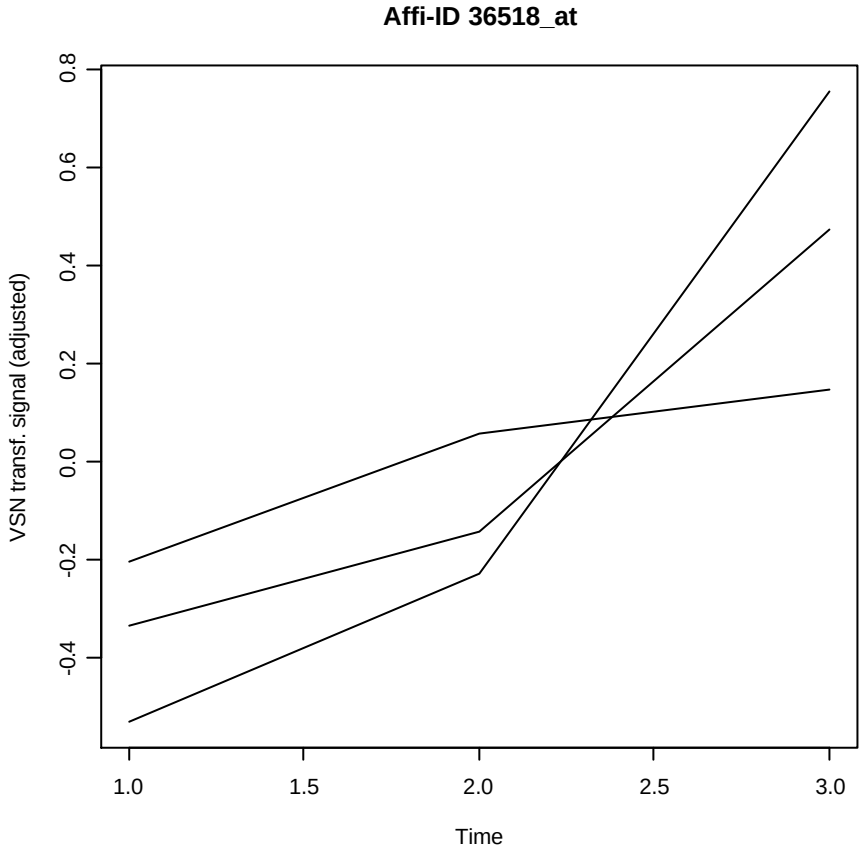
Analysis of variance(ANOVA) studies differences between then mean values of normal distributed data from groups.

We are interested in an analysis of $y_{ij} = x_{ij} - \bar{x}_i$, time course measurements adjusted for individual probe levels

The biological variability between individual probes is not of interest for the next step and is therefore separated.

We only look at *individually adjusted* time courses.

Important boundary condition: $\sum_{j=1}^{3} y_{ij} = 0$ for all i (i = 1, 2, 3)

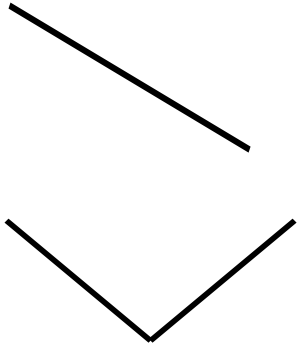


Is there evidence for a *systematic* pattern?
What is a *systematic* pattern?

ANOVA – Analysis of variance (2)

A formal view on the data of patient i (probe i)

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix} = a \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ (-1) \cdot \frac{1}{\sqrt{2}} \end{pmatrix} + b \cdot \begin{pmatrix} \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \end{pmatrix}$$



A sequence of three measurements with mean 0 can be decomposed into a *linear* term, a quadratic term, and residuals.

How to calculate the coefficients a and b?

ANOVA – Analysis of variance (3)

y_{11}	=	a	$2^{-1/2}$	+	b	$6^{-1/2}$	$+ \varepsilon_{11}$
y_{12}	=				b	$(-2)\cdot 6^{-1/2}$	$+ \varepsilon_{12}$
y_{13}	=	a	$(-2^{-1/2})$	+	b	$6^{-1/2}$	$+ \varepsilon_{13}$
y_{21}	=	a	$2^{-1/2}$	+	b	$6^{-1/2}$	$+ \varepsilon_{21}$
y_{22}	=				b	$(-2)\cdot 6^{-1/2}$	$+ \varepsilon_{22}$
y_{23}	=	a	$(-2^{-1/2})$	+	b	$6^{-1/2}$	$+ \varepsilon_{23}$
y_{31}	=	a	$2^{-1/2}$	+	b	$6^{-1/2}$	$+ \varepsilon_{31}$
y_{32}	=				b	$(-2)\cdot 6^{-1/2}$	$+ \varepsilon_{32}$
y_{33}	=	a	$(-2^{-1/2})$	+	b	$6^{-1/2}$	$+ \varepsilon_{33}$

Least square estimates for coefficients:

$$\bar{a} = \frac{\sqrt{2}}{3 \cdot 2} \sum_{i=1}^{3} (y_{i1} - y_{i3}) \quad \text{and} \quad \bar{b} = \frac{\sqrt{6}}{3 \cdot 4 \cdot} \sum_{i=1}^{3} (y_{i1} + y_{i3} - y_{i2})$$

Residuals ε are calculated by taking the difference between observed and model based values.

ANOVA – Analysis of variance (4)

The final decomposition of the variance

$$SS = \sum_{i=1}^{3}\sum_{j=1}^{3}(x_{ij} - \bar{x})^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}(x_{ij} - \bar{x}_i)^2 + 3\sum_{i=1}^{3}(\bar{x}_i - \bar{x})^2 =$$

$$\sum_{i=1}^{3}\sum_{j=1}^{3}(\varepsilon_{ij})^2 + \qquad 3\bar{a}^2 + 3\bar{b}^2 + \qquad 3\sum_{i=1}^{3}(\bar{x}_i - \bar{x})^2$$

unexplained variability	model based variability	patient level variability

ANOVA – Analysis of variance (5)

```
> breit.vsn[6,]
   P.1.0    P.1.1    P.1.2    P.2.0    P.2.1    P.2.2    P.3.0    P.3.1    P.3.2
14.17067 14.36492 14.97927 14.01884 14.28024 14.37200 13.79599 14.09769 15.08175

> contrasts(ordered(1:3))
                    .L          .Q
       1 -7.071068e-01  0.4082483
       2 -7.850462e-17 -0.8164966
       3  7.071068e-01  0.4082483

> breit.simple.anova.rfc
function (gene.nr=6)
       {
              cc<-contrasts(ordered(1:3))
              yy<-breit.vsn[gene.nr,]
              pp<-as.factor(rep(1:3,c(3,3,3)))
              ll<-rep(cc[,1],3)
              qq<-rep(cc[,2],3)
              return(aov(yy~ll+qq+Error(pp)))
       }
```

ANOVA – Analysis of variance (6)

Results for gene *36518_at*

```
> summary(breit.simple.anova.rfc())

Error: pp
          Df   Sum Sq  Mean Sq F value Pr(>F)
Residuals  2 0.121730 0.060865

Error: Within
          Df  Sum Sq Mean Sq F value Pr(>F)
ll         1 0.99838 0.99838 14.2146 0.0196 *
gg         1 0.04834 0.04834  0.6883 0.4534
Residuals  4 0.28095 0.07024
```

$$SS = \sum_{i=1}^{3}\sum_{j=1}^{3}(\varepsilon_{ij})^2 + 3\bar{a}^2 + 3\bar{b}^2 + 3\sum_{i=1}^{3}(\bar{x}_i - \bar{x})^2$$

All genes together

global and *gene specific* trends

$x_{k,i,j}$ = Signal of gene k in patient (probe) i at time j

$$x_{k,i,j} = m_{k,i} + l(j) + q(j) + l_k(j) + q_k(j) + \varepsilon_{k,i,j}$$

global (gene independent) linear effect:	$l(1) = a / \sqrt{2}$,	$l(2) = 0$,	$l(3) = - l(1)$
gene specific linear effect:	$l_k(1) = a_k / \sqrt{2}$,	$l(2) = 0$,	$l(3) = - l_k(1)$
global (gene independent) quadratic effect:	$q(1) = b / \sqrt{6}$,	$q(2) = -2\ q(1)$,	$q(3) = q(1)$
gene specific quadratic effect:	$q_k(1) = b_k / \sqrt{6}$,	$q_k(2) = -2\ q_k(1)$	$q_k(3) = q_k(1)$

important restrictions:

$$\sum_{k=1}^{K} l_k(j) = \sum_{j=1}^{3} l_k(j) = \sum_{k=1}^{K} q_k(j) = \sum_{j=1}^{3} q_k(j) = 0$$

ANOVA – Analysis of variance (7)

Estimation of coefficients

$$a = \frac{\sqrt{2}}{3 \cdot 2 \cdot K} \sum_{k=1}^{K} \sum_{i=1}^{3} (x_{ki1} - x_{ki3}) \qquad b = \frac{\sqrt{6}}{3 \cdot 4 \cdot K} \sum_{k=1}^{K} \sum_{i=1}^{3} (x_{ki1} + x_{ki3} - 2x_{ki2})$$

$$a_k = \frac{\sqrt{2}}{3 \cdot 2} \sum_{i=1}^{3} (x_{ki1} - x_{ki3}) - a \qquad b_k = \frac{\sqrt{6}}{3 \cdot 4} \sum_{i=1}^{3} (x_{ki1} + x_{ki3} - x_{ki2}) - b$$

$$e_{k,i,j} = (x_{k,i,j} - x_{k,i,\cdot}) - [l(j)+q(j)+l_k(j)+ q_k(j)] \qquad x_{k,i,\cdot} = (x_{k,i,1} + x_{k,i,2} + x_{k,i,3}) / 3$$

$$\text{within group variability: } \frac{1}{2 \cdot 2 \cdot K} \sum_{k=1}^{K} \sum_{i=1}^{3} \sum_{j=1}^{3} e_{k,i,j}^2$$

$$\text{between patient (probe) variability: } \frac{1}{2 \cdot K} \sum_{k=1}^{K} \sum_{i=1}^{3} (x_{k,i,\cdot} - x_{k,\cdot,\cdot})^2$$

$x_{k,\cdot,\cdot}$ Mean of gene k

ANOVA – Analysis of variance (8)

Results and ANOVA table

```
> breit.complex.anova.rfc()
                    DF            SS           MS            FF
Between          25250 4.723376e+03  0.18706440            NA
linear.global        1 2.225109e+01 22.25109107 144.52207894
quadr.global         1 1.093243e-02  0.01093243   0.07100672
linear.gene      12624 7.081756e+03  0.56097558   3.64356770
quadr.gene       12624 4.493544e+03  0.35595247   2.31193119
Within           50500 7.775145e+03  0.15396326            NA
```

There are 9 x 12625 = 113625 measurement
12625 DF are lost because of block approach with respect to each gene per patient.
The within part therefore accounts for 113625 – 12625 = 101000 DF.

```
25250 + 1 + 1 + 12624 + 12624 + 50500 = 101000
```

Questions of interest (1)

1. Is there a global linear effect (slightly decreasing a ~ 0.024)?
2. Are there relevant gene-specific effects:

	# genes < (-1)	# genes > 1
a.k	153	318
b.k	56	82

3. Which genes have relevant gene specific effects?
4. Are the *between probe* and the *within probe* variability similar?
5. Residuals are not normally distributed

More questions?

Questions of interest (2)

Is there a global linear effect ?

Answer:

1. Test the null hypothesis *a=0*

2. If the null is rejected, calculate a 95% confidence interval for a and assess based on the confidence interval if the effect is *relevant*.

The linear effect was estimated as a=0.024.
If this effect could be extrapolated to longer time intervals,
it would take 30 days to half the expression level measured at time 0.
30 x 0.024 ~ log(2)

Questions of interest (3)

Are there relevant gene-specific effects?
Which genes have relevant gene specific effects?

Answer:

1. Test the null hypothesis $a_k = b_k = 0$

2. If the null is rejected, calculate an appropriate confidence interval for a_k and b_k (k = 1, ..., K)

3. Look for genes where the confidence intervals do not contain 0.

Questions of interest (4)

Are the *between probe* and the *within probe* variability similar?

Answer:

Based on the appropriate model calculate a 95% confidence interval for the difference or quotient of the *within* and *between* variability.

The big problem

The distribution of the residuals has heavier tails as the normal distribution.

Therefore, the classical ANOVA theory can not be used to answer questions of inferential statistics concerning

1. The test of null hypotheses?
2. Calculation of confidence intervals?

How can we perform statistical tests and calculate confidence intervals without knowing the parametric form of the relevant distributions?

The Bootstrap – basic idea I

1. Imitate an unknown random process based on the observed data.
2. Transfer the results gained by imitating the unknown random process to the unknown random process.

It is truly important that the distribution of the *imitation* is close to the unknown but true distribution.

The Bootstrap – basic idea II

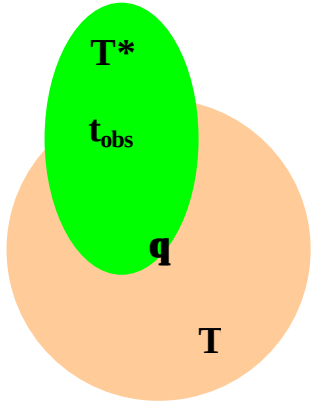
T random variable which describes the statistic (parameter) under the true, but unknown distribution.

T* random variable which describes the statistic under the bootstrap distribution.

Evaluation of $T-\theta$ by T^*-t_{obs} has two sources of error:

- data variability between true and bootstrap distribution.
- estimates based on finite simulations (simulation error).

(θ - parameter to be estimated, t_{obs} – estimate of θ from the data)



The Bootstrap – testing a null-hypothesis

Question of interest:	Are the gene specific components important to explain the variability observed in our experiment?
Alternative:	M_A: $l+q+l_k+q_k$
Null-hypothesis:	M_{Ho}: The true model is given by $l+q$
Test statistic :	F-value for the l_k and q_k components if M_A is fitted to the data. ($F_{lk,obs}$ = 3.64 and $F_{qk,obs}$ = 2.31)
Distribution of T*:	Fit M_{Ho} to the observed data and calculate residuals and fitted values. *Create a new data* set by resampling with replacement from the residuals. Add the new residuals to the fitted values. *Calculate t** by fitting Model M_A to the new data set Repeat both steps many times (1000 times)
Bootstrap p – value:	$p_{boot} = \#\{t^* \geq t_{obs}\}$ / # bootstrap samples (p_{boot} = 0 for l_k and q_k)

The Bootstrap – confidence intervals (1)

Quantiles of T-θ will be approximated by using ordered values of T*-t_{obs}
(t* is a realisation of T*, bootstrap sample t_1^*, ..., t_N^*)

Estimate p quantile of T-θ by the $(N+1)\cdot p^{th}$ ordered value of the bootstrap sample
$\{t_1^*-t_{obs}, ..., t_N^*-t_{obs}\}$, that is $t^*_{[(N+1)\cdot p]} - t_{obs}$ (the value of $(N+1)\cdot p$ has to be an integer)

Simple (1-α) confidence intervals for θ:

Quantile method: $[t_{obs} - (t^*_{[(N+1)\cdot(1-\alpha/2)]}-t_{obs}); t_{obs} - (t^*_{[(N+1)\cdot\alpha/2]}-t_{obs})]$

Studentized: $[t_{obs} - z_{1-\alpha/2}\cdot\sqrt{v^*} ; t_{obs} + z_{1-\alpha/2}\cdot\sqrt{v^*}]$
v^* is bootstrap estimate of variance of the mean of T*.

The Bootstrap – confidence intervals (2)

Question of interest: Which genes in our experiment of show gene-specific effects?
The argument will be based on the 99% confidence interval for the contrast a_k and b_k.

- There are simultaneously 12625 statistics of interest, one for each.
- What is an appropriate procedure to sample realisations t_k^* of T_k^*? [See Wu CFJ (1986)]

  - *Fit the model* M_A : $\mu+l+q+l_k+q_k$ to the data. Get the fitted values and the residuals.
  - *Create a new data set* by resampling with replacement from the residuals. Rescale the sampled residuals by $[6xK/(K-6)]^{1/2}$. Add the rescaled residuals to the fitted values.
  - *Calculate a new realisation of* T_k^*, k = 1, ..., K.
  - *Repeat the last two steps many times* (20000 times)
  - For each k = 1, ..., K calculate a (1-α) *bootstrap confidence interval* (α = 0.01).
- Multiple testing problems?

The Bootstrap – Caveats

- Quantiles depend on the sample in an unsmooth or unstable way. For finite samples it may not work well. The set of possible values for T* may be very small and vulnerable to unusual data points.

- Incomplete data: The missing mechanism has to be non-informative to guarantee the statistical consistency of the estimation of T.

- Dependent data: Bootstrap estimate of the variance would be wrong.

- Dirty data: Outliers in the data may imply that the conclusions depend crucially on particular observations (especially in the non-parametric case).

References

1. Kerr MK, Martin M, Churchill GA (2000), *Analysis of Variance for Gene Expression Microarray Data*, Journal of Computational Biology, 7: 819-837.
2. Wu CFJ (1986) *Jackknife, Bootstrap, and other Resampling Methods in Regression Analysis,* Annals of Statistics, 14: 1261-1295.
3. Davison AC, Hinkley DV (1998) *Bootstrap methods and their application*, Cambridge University Press, Cambridge.
4. Tusher VG, Tibshirani R, Chu G. (2001) *Significance analysis of microarrays applied to the ionizing radiation response,* PNAS, 98: 5116-5121.