Resampling and the Bootstrap

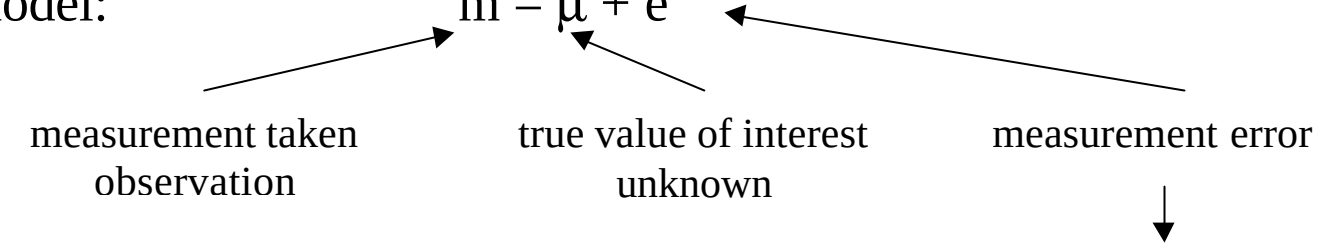
Ulrich Mansmann

mansmann@imbi.uni-heidelberg.de

Practical microarray analysis
January 2003
Berlin

Motivation

- Measurement model: $m = \mu + e$



- The great challenge is, how to gain information on the error distribution.

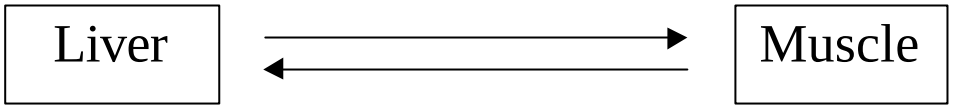
For doing statistics the measurement error has to fulfil the following criteria:

- $E[e] = 0$
- $Var[e] = \sigma^2$ – variation independent of μ.
- To calculate CI's and to perform statistical tests the distribution of e has to be known.

It can not generally be assumed that $e \sim N(0, \sigma^2)$

Motivation – Example (1)

Analysis of Variance for Gene Expression Microarray Data [Kerr et al. (2000)]



Which genes are differentially expressed?

Model: $m_{i,j,k,g} = \mu + A_i + D_j + V_k + G_g + AG_{ig} + VG_{kg} + \varepsilon_{i,j,k,g}$

	Array	
Dye	1	2
1	Liver (1,1,1)	Muscle (2,1,2)
2	Muscle (1,2,2)	Liver (2,2,1)

$(i,j,k) \in \{ (1,1,1), (1,2,2), (2,1,2), (2,2,1)\}$

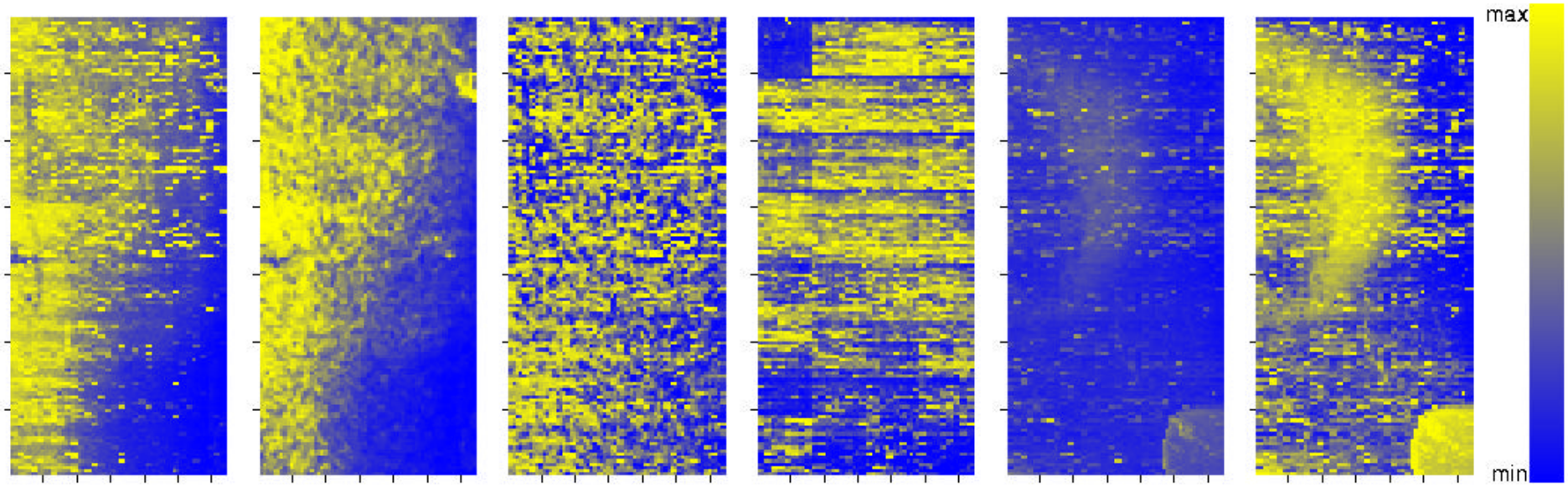
AG: Array – Gene Interaction

VG: Variety – Gene Interaction

Interest: For which genes is $VG_{1g} - VG_{2g} \neq 0$?

$\varepsilon_{i,j,k,g}$ is independent – co-regulation of genes is hidden in G_g and VG_{kg}

ANOVA models: $(AG)_{ig}$



W. Huber, DKFZ, Heidelberg

Motivation – Example (2)

Estimation of model parameters

$\mu = \text{mean}(m_{ijkg})$

$A_1 = \text{mean}(m_{1jkg}) - \mu$; $A_2 = \text{mean}(m_{2jkg}) - \mu$
$D_1 = \text{mean}(m_{i1kg}) - \mu$; $D_2 = \text{mean}(m_{i2kg}) - \mu$
$V_1 = \text{mean}(m_{ij1g}) - \mu$; $V_2 = \text{mean}(m_{ij2g}) - \mu$
$G_{g*} = \text{mean}(m_{ijkg*}) - \mu$

$AG_{i*g*} = \text{mean}(m_{i*jkg*}) - A_{i*} - G_{g*} - \mu$
$VG_{k*g*} = \text{mean}(m_{ijk*g*}) - V_{k*} - G_{g*} - \mu$

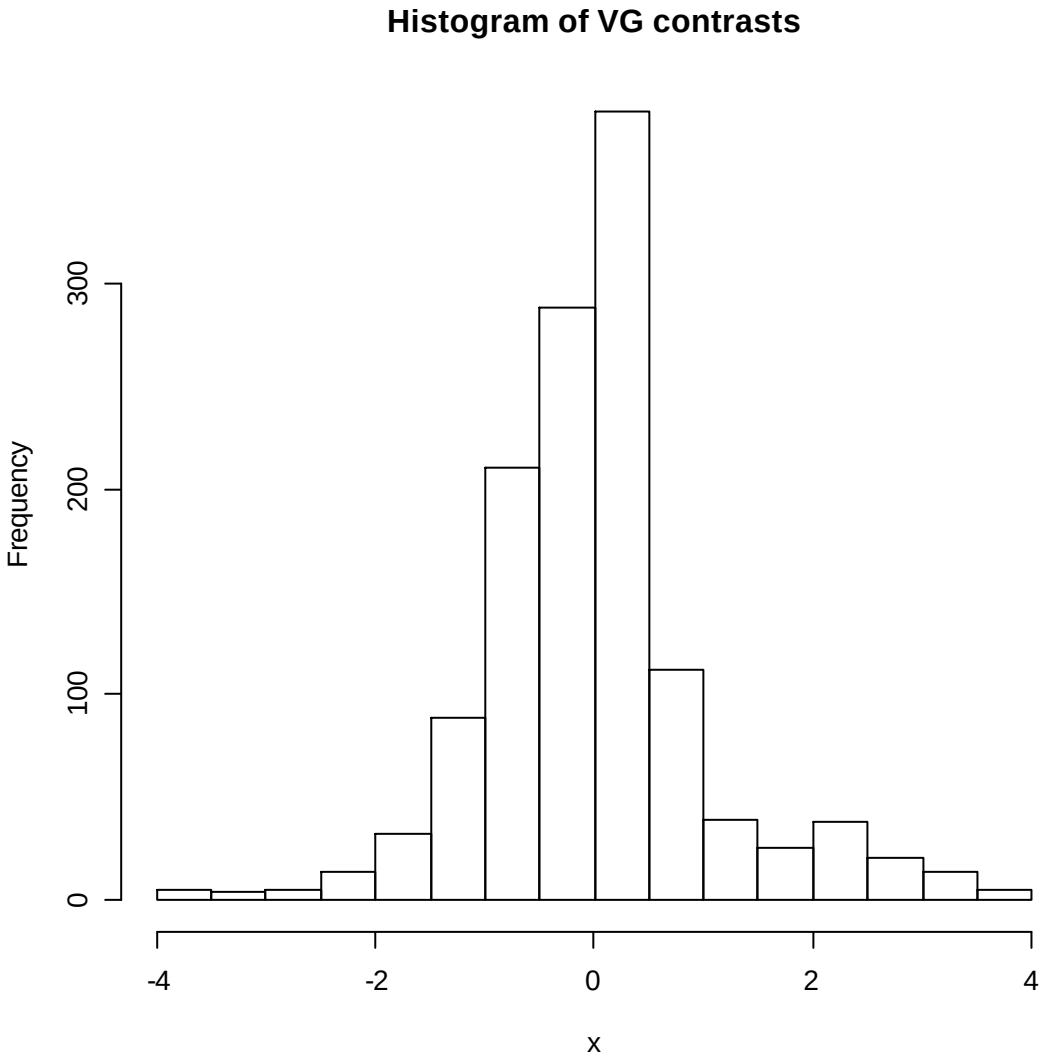
$A_1+A_2 = \mu_{1\ldots} + \mu_{2\ldots} - 2\mu = 0$
$D_1+D_2 = \mu_{.1..} + \mu_{.2..} - 2\mu = 0$
$V_1+V_2 = \mu_{..1.} + \mu_{..2.} - 2\mu = 0$
$\Sigma G_g = 0$

$\Sigma_i AG_{ig} = \Sigma_g AG_{ig} = 0$
$\Sigma_k VG_{kg} = \Sigma_g VG_{kg} = 0$

	Array	
Dye	1	2
1	Liver	Muscle
2	Muscle	Liver

i*, k*, g* - fixed indices

Motivation – Example (3)




```
ls.figure.1.rfc()
```

Motivation – Example (4)

ANOVA results

```
> ls.effect.estimates.res <- ls.effect.estimates.rfc()
> ls.effect.estimates.res$ANOVA.table
```

	DF	SS	MS	F
Array	1	92.338	92.338	1433.870
Dye	1	0.744	0.744	11.558
Variety	1	2.969	2.969	46.106
Gene	1285	1885.893	1.468	22.790
ArrayxGene	1285	160.014	0.125	1.934
VarietyxGene	1285	1357.283	1.056	16.402
Residual	1285	82.751	0.064	
Total	5143	3581.991		

random variation

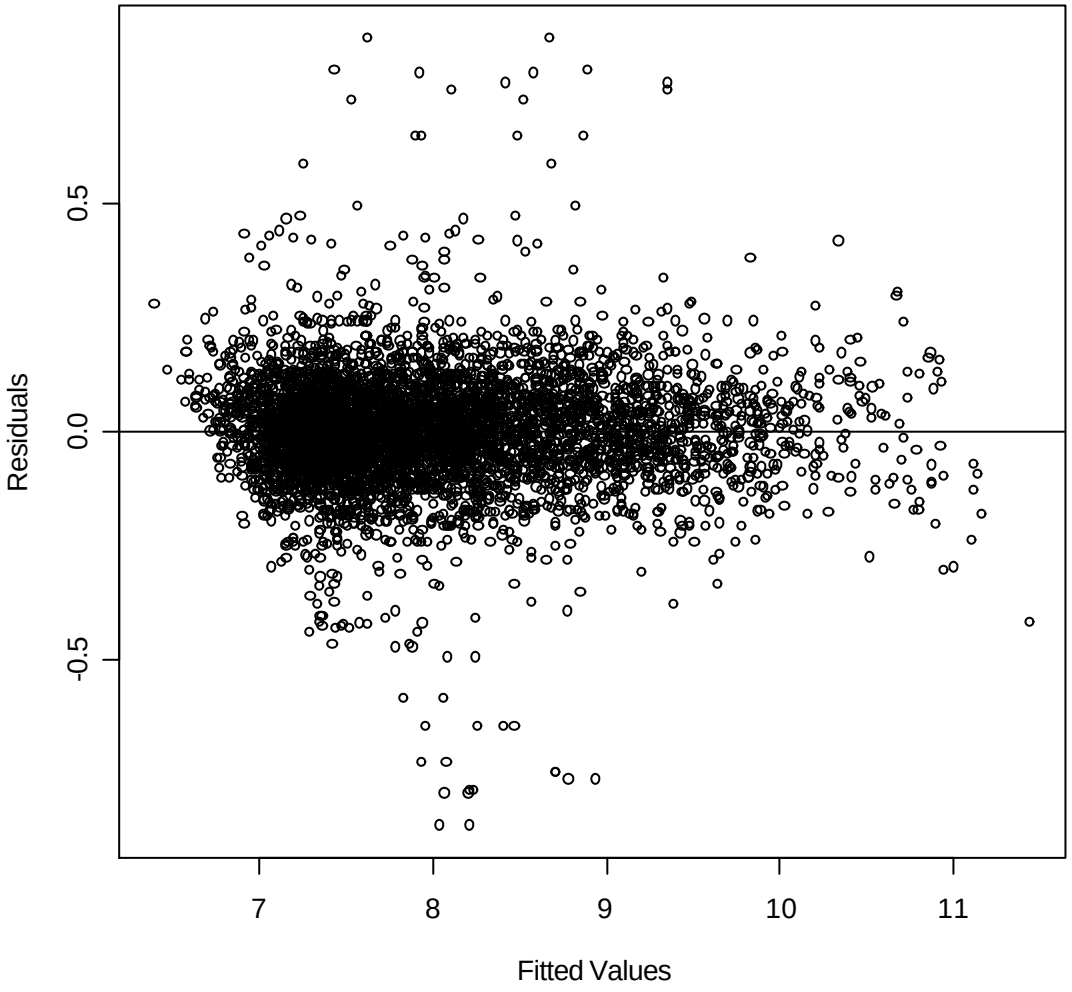
$$SS_{total} = \sum (m_{ijkg} - \mu)^2 = \sum_{i,j,k,g} resid_{ijkg}^2 +$$

$$\sum_{i,j,k,g} A_i^2 + \sum_{i,j,k,g} D_j^2 + \sum_{i,j,k,g} V_i^2 + \sum_{i,j,k,g} G_g^2 + \sum_{i,j,k,g} AG_{ig}^2 + \sum_{i,j,k,g} VG_{kg}^2$$

Variation explained by the model

Motivation – Example (5)

Measurement model



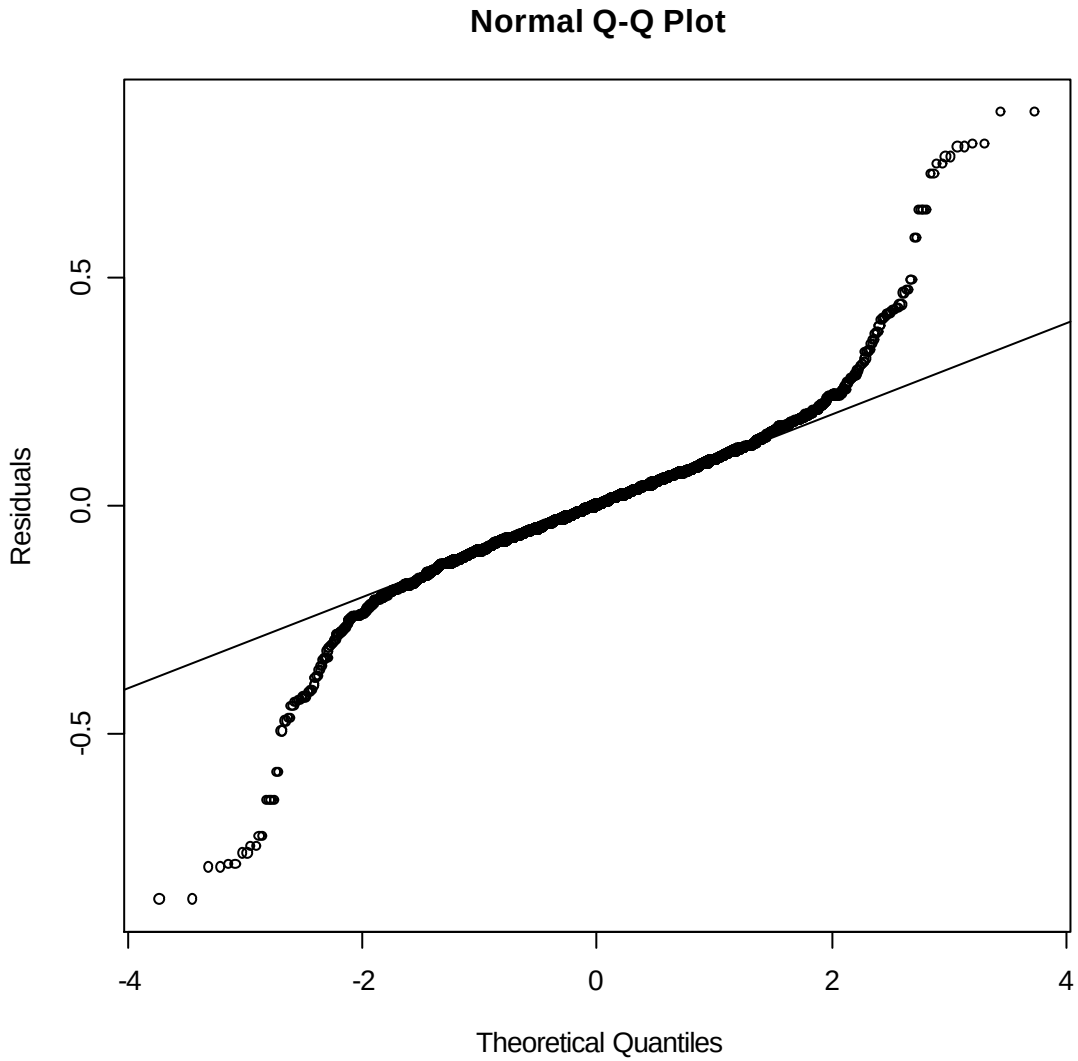

```
ls.figure.3.rfc()
```

$\sigma^2 = 0.02$

$\sigma = 0.13$

Motivation – Example (6)

Measurement model




```
ls.figure.2.rfc()
```

Motivation – Example (7)

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 like:

1. Is the AG component (array-gene interaction) an important component of the model (F=1.934)?
2. Which genes are differential expressed?

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.

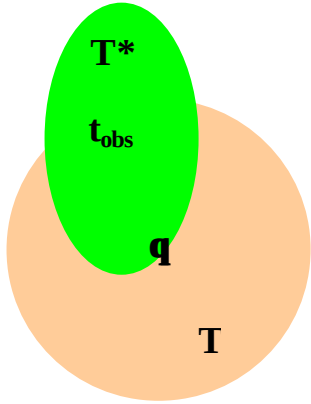
The Bootstrap – basic idea II

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-θ 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)

Example: CI for the mean of a Normal distribution.

```
x<-rnorm(20,0,1)
```

```
m = mean(x) = -0.2739045
v = var(x)  =  1.230081, s = std(x)  =  1.109090
```

We imitate the random process: we assume that the data is produced from a normal distribution with mean = -0.2739045 and var = 1.230081

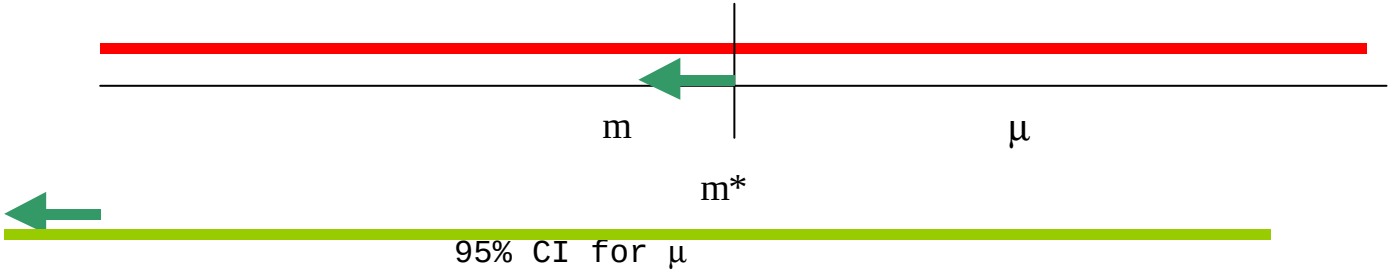
Now we perform (say 999) studies with 20 patient from this random mechanism

```
> x.star<-NULL
> for (i in 1:999) x.star<-c(x.star,mean(rnorm(20,-0.2739045,1.109090)))
```

and get the following result:

```
> m* = mean(x.star) = -0.2663
> quantile(x.star,c(0.025,0.975)):          2.5%      97.5%
                                    -0.7588371  0.2004343
```

95% CI for m



The Bootstrap – two approaches

- Parametric bootstrap (PB):
  Mathematical model for the distribution of interest is known and has parameter ψ. The parameter to be estimated θ is a component of ψ.
  Example: Normal distribution $\psi = (\mu, \sigma^2)$, $\theta = \mu$.

- Non-parametric bootstrap (NPB):
  No mathematical model available. Approach is based on the fact that data is based on observations from ***iid*** (independent identically distributed) random variables.
  NPB may also be used to assess the robustness of conclusions drawn from a parametric analysis.

```
Example (95% CI for μ):

Parametric bootstrap: [-0.7664; 0.1928]
Wild bootstrap:       [-0.7116  0.1788]
Exact procedure:      [-0.7930; 0.2452]
```

The Bootstrap – testing a null-hypothesis (1)

Question of interest: Is the AG-component important to explain the variability observed in the Kerr et al. experiment?

Alternative: M_A: μ+A+D+V+G+AG+VG

Null-hypothesis: M_{Ho}: The true model is given by μ+A+D+V+G+VG

Test statistic T: F-value for the AG component if M_A is fitted to the data. (t_{obs} = 1.934)

Distribution of T*: Fit M_{Ho} to the observed data and calculate residuals and fitted values.
Under the null-hypothesis, the 5144 residuals are iid.
Create a new data set by resampling with replacement 5144 times 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 (say 1000 times)
Result: Min = 0.852, Max = 1.254 (1000 samples)
Min = 0.81, Max = 1.27 (20000 samples, Kerr et al.)

Bootstrap p – value: $p_{boot} = \#\{t^* \geq t_{obs}\} / \#$ bootstrap samples (= 0)

Procedure is given programmed in `ls.boot.F.value.rfc`.

The Bootstrap – testing a null-hypothesis (2)

- **Improving the bootstrap p-value** and checking if bootstrap test is ok:
  $p_{boot} = P(T^* \geq t_{obs} \mid \text{obs. Data})$ has to be uniformly distributed.
  Can be checked by a double bootstrap.

  Improved bootstrap p-value: $p_{adj} = P(P^* \leq p_{boot} \mid \text{obs. Data})$
  P^* is a random variable which describes the behaviour of p_{boot} under a double bootstrap.

- **Estimating the power of a test**:
  Power is defined by $P(T \geq t_p \mid H_A)$ where t_p comes from $P(T \geq t_p \mid H_0) = p$

  The $t_{0.05}$ in the Kerr example can be estimated (1000 samples) by 1.096. The bootstrap algorithm proposed for the test problem can be easily modified to sample data under the alternative H_A for some prespecified AG effect.

The Bootstrap – confidence intervals (1)

Quantiles of $T-\theta$ 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 $T-\theta$ 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 $(1-\alpha)$ 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 the experiment of Kerr et al. are differentially expressed? Kerr et al. base the argument on the 99% confidence interval for the contrast VG_{1g}-VG_{2g}.

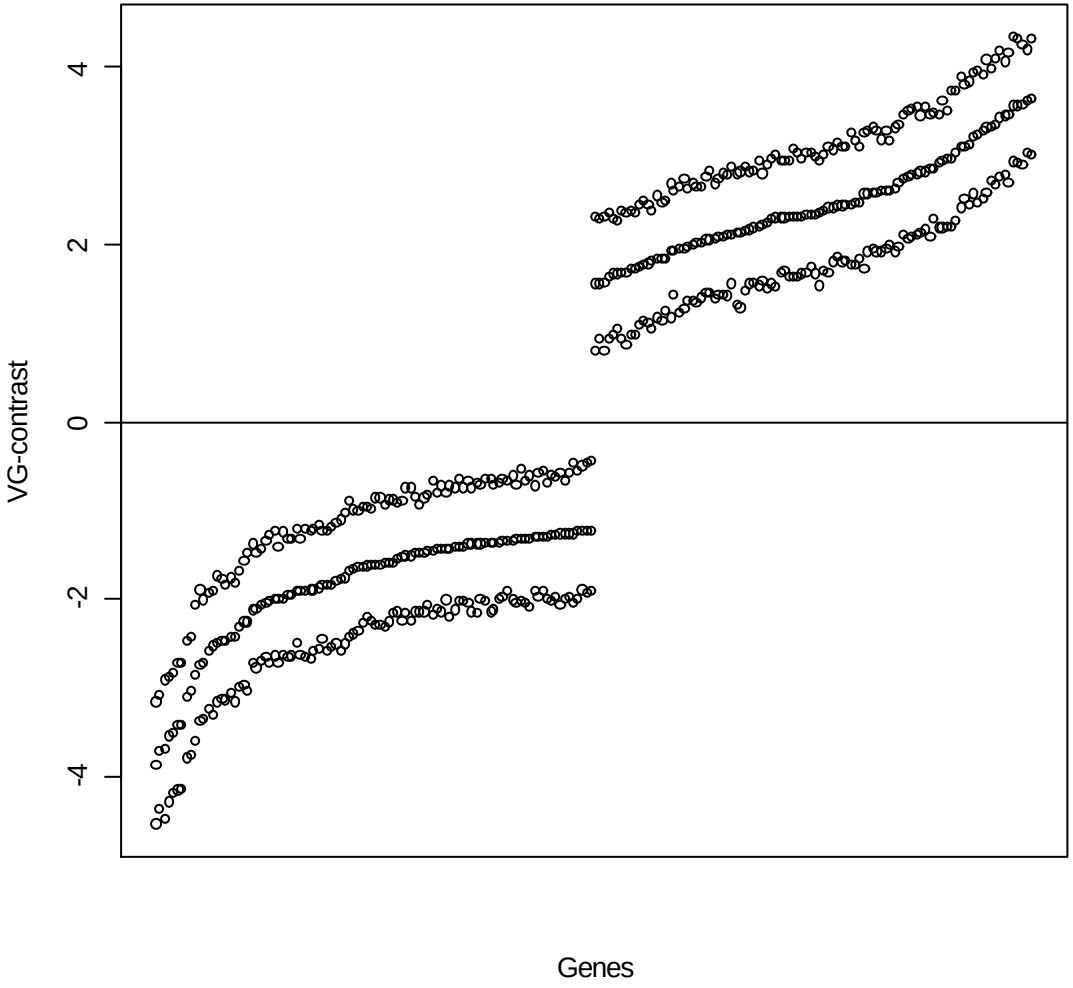
- There are simultaneously 1286 statistics of interest, one for each gene: $T_g^* = VG_{1g}-VG_{2g}$.
- What is an appropriate procedure to sample realisations t_g^* of T_g^*? [See Wu CFJ (1986)]

  - *Fit the model* M_A : μ+A+D+V+G+AG+VG to the data. Get the fitted values and the residuals.
  - *Create a new data set* by resampling 4x1286 with replacement from the residuals. Rescale the samples residuals by $[4xN/(N-4)]^{1/2}$. Add the rescaled residuals to the fitted values.
  - *Calculate a new realisation of* T_g^*, g = 1, ..., N.
  - *Repeat the last two steps many times* (Kerr et al. 20000 times)
  - For each g = 1, ..., N calculate a (1-α) *bootstrap confidence interval* (α = 0.01).

  Procedure programmed in `ls.boot.expression.rfc`.

- Multiple testing problems?

The Bootstrap – confidence intervals (3)



Genes are ordered with respect to the size of the contrast $VG_{1g} - VG_{2g}$. The 100 genes with the smallest and the 100 genes with the highest contrast values are plotted together with a 99% bootstrap confidence intervals.

The mean width of the CIs is 1.384 (5000 samples).

A exp{1.384/2} = 1.998 fold change in gene expression implies a systematic differential expression.

`ls.figure.4.rfc`

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).

The Permutation Test (1)

- The permutation test is a test where the null-hypothesis allows to reduce the inference to a randomisation problem. The process of randomisation makes it possible to ascribe a probability distribution to the difference in the outcome possible under H_0.
- The outcome data are analysed many times (once for each acceptable assignment that could have been possible under H_0) and then compared with the observed result, without dependence on additional distributional or model-based assumptions.
- How to perform a permutation test:
  - Analysis of the problem, choice of null-hypothesis
  - Choice of test statistic T
  - Calculate the value of the test statistic for the observed data: t_{obs}
  - Apply the randomisation principle and look at all possible permutations, this gives the distribution of the test statistic T under H_0.
  - Calculation of p-value: $p = P(T \geq t_{obs} \mid H_0) \sim \#\{t^* \geq t_{obs}\} / \#$ permutations.

The Permutation Test (2)

Coexpression of genes

Data: n measurements of gene expression for Gene 1 and Gene 2 with respect to some reference – (g_1^1, g_1^2), ..., (g_n^1, g_n^2).

H_0: Gene 1 and Gene 2 are not correlated.

Test statistic T: Pearson (or Spearman) correlation coefficient, calculate t_{obs}

Randomization: $(g_{(1)}^1, ..., g_{(n)}^1, g_{(1)}^2, ..., g_{(n)}^2)$

Under H_0 it is possible to permute the values observed for Gene 2. There are n! possibilities.

⇒ Distribution of T under H_0

p-value: p = # of permutations such that $T^* \geq t_{obs}$ / n!

The Permutation Test (3)

- A comparative test between two groups of sizes n_1 and n_2 has to look at $\binom{n_1+n_2}{n_1}$ permutations of the group indices.

  CAVE: What is H_0? For the test the H_0: $F_1 = F_2$ is needed. This does not follow from H_0: $\mu_1 = \mu_2$. Additional assumptions are needed: $F_i(y)= G(y-\mu_i)$ or $= G(y/\mu_i)$ for G unspecified.

- Special example: SAM [Tusher et al. (2001)]

- It is rarely possible or necessary to compute the permutation p-value exactly. The most practical approach is to take a large number N of random permutations, calculate the corresponding values $t_1^*, ..., t_N^*$ of T, and approximate p by

$$p_{mc}= [1 + \#\{t^* \geq t_{obs} \quad \}]/(N+1)$$

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.