Recap: General Linear Model

Lieven Clement

Ghent University, Belgium

Statistical Genomics: MSc in Bioinformatics & MSc in Statistical Data Analysis

Breast cancer example (part of study https://doi.org/10.1093/jnci/djj052)

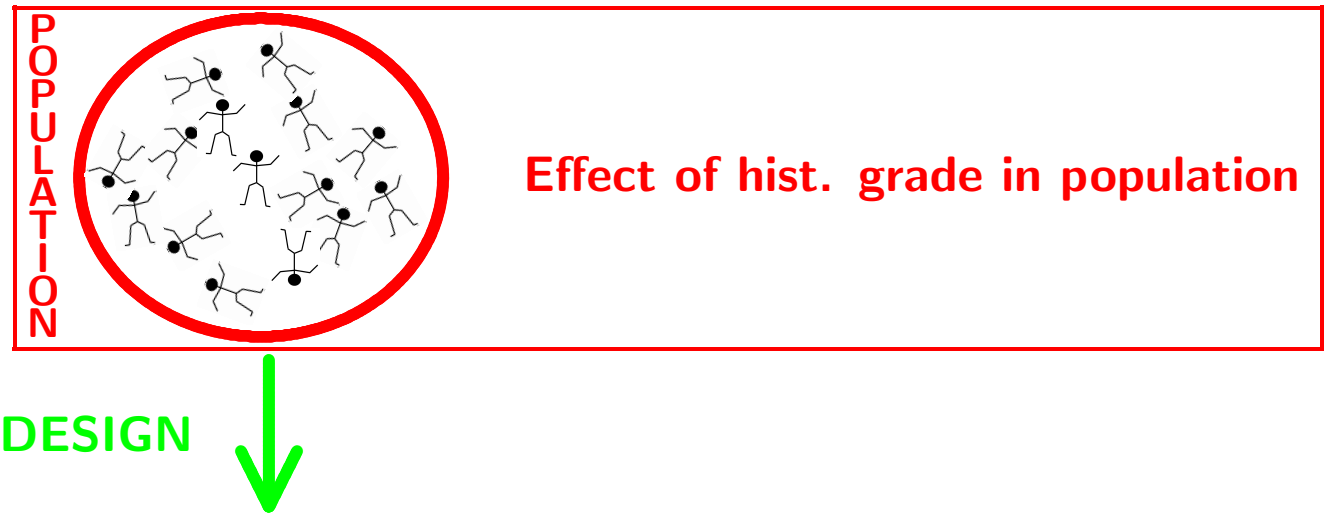
- Histologic grade in breast cancer clinically prognostic. Association of histologic grade on expression of KPNA2 gene that is known to be associated with poor BC prognosis.

Breast cancer example (part of study https://doi.org/10.1093/jnci/djj052)

- Histologic grade in breast cancer clinically prognostic. Association of histologic grade on expression of KPNA2 gene that is known to be associated with poor BC prognosis.
- Population: all current and future breast cancer patients

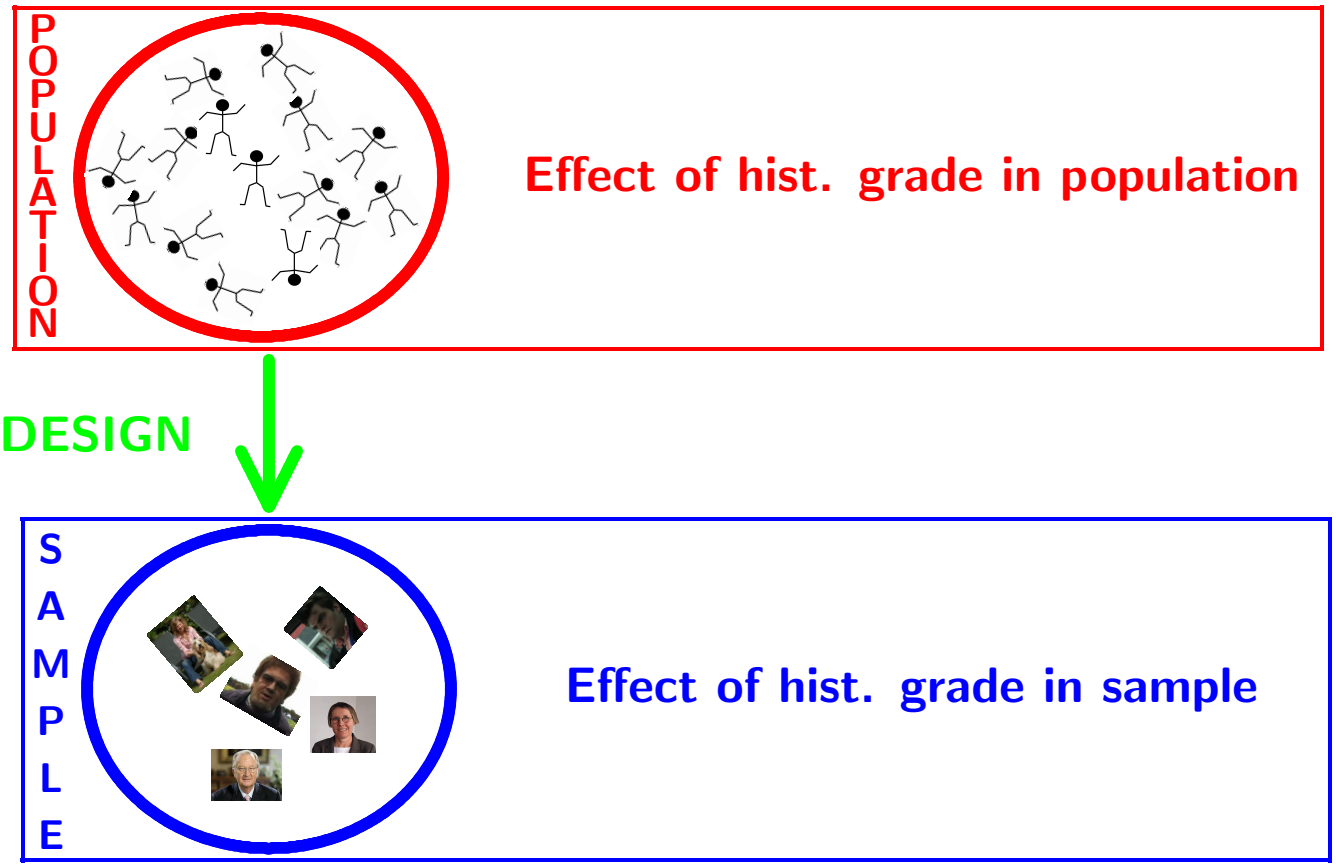
Breast cancer example (part of study https://doi.org/10.1093/jnci/djj052)

- Histologic grade in breast cancer clinically prognostic. Association of histologic grade on expression of KPNA2 gene that is known to be associated with poor BC prognosis.
- Population: all current and future breast cancer patients



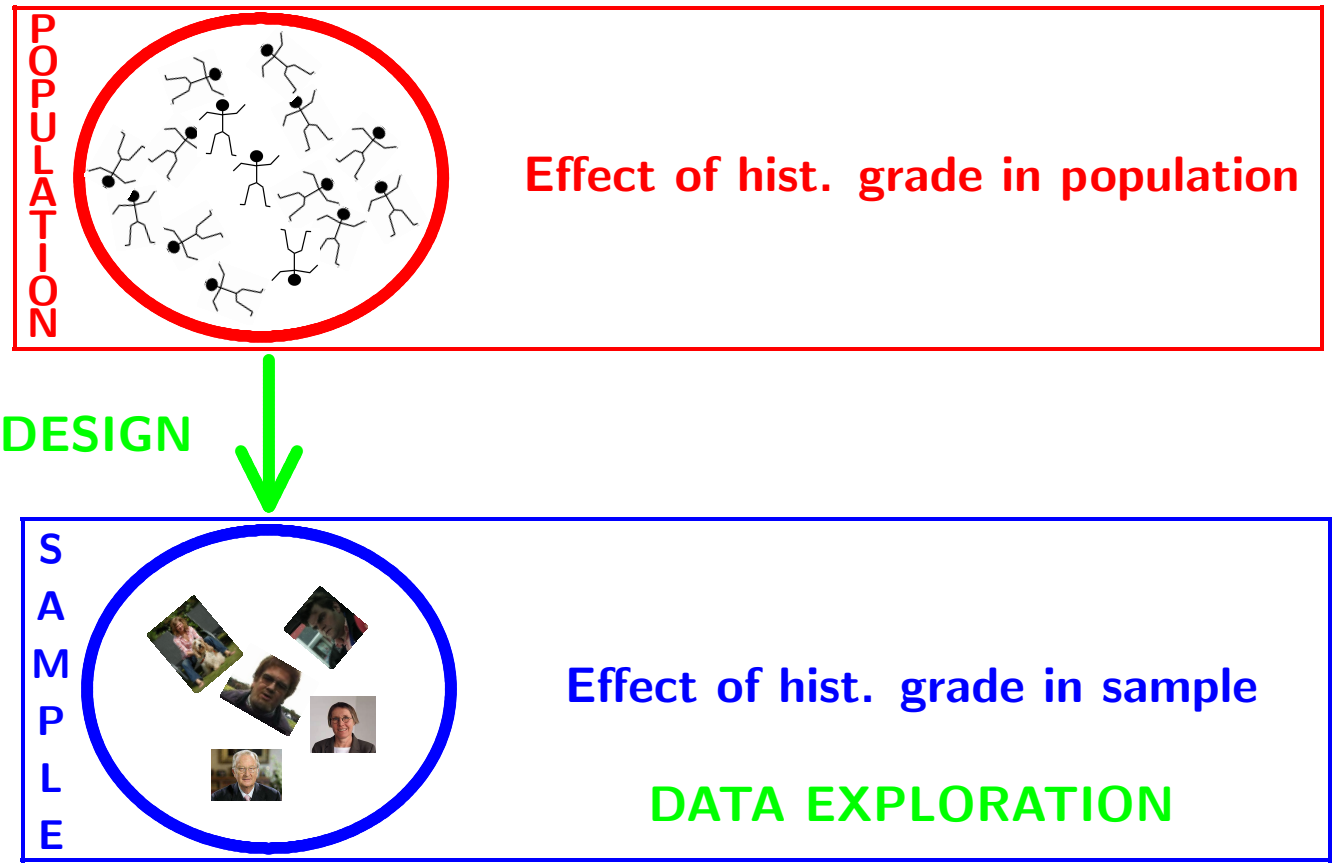
Breast cancer example (part of study https://doi.org/10.1093/jnci/djj052)

- Histologic grade in breast cancer clinically prognostic. Association of histologic grade on expression of KPNA2 gene that is known to be associated with poor BC prognosis.
- Population: all current and future breast cancer patients



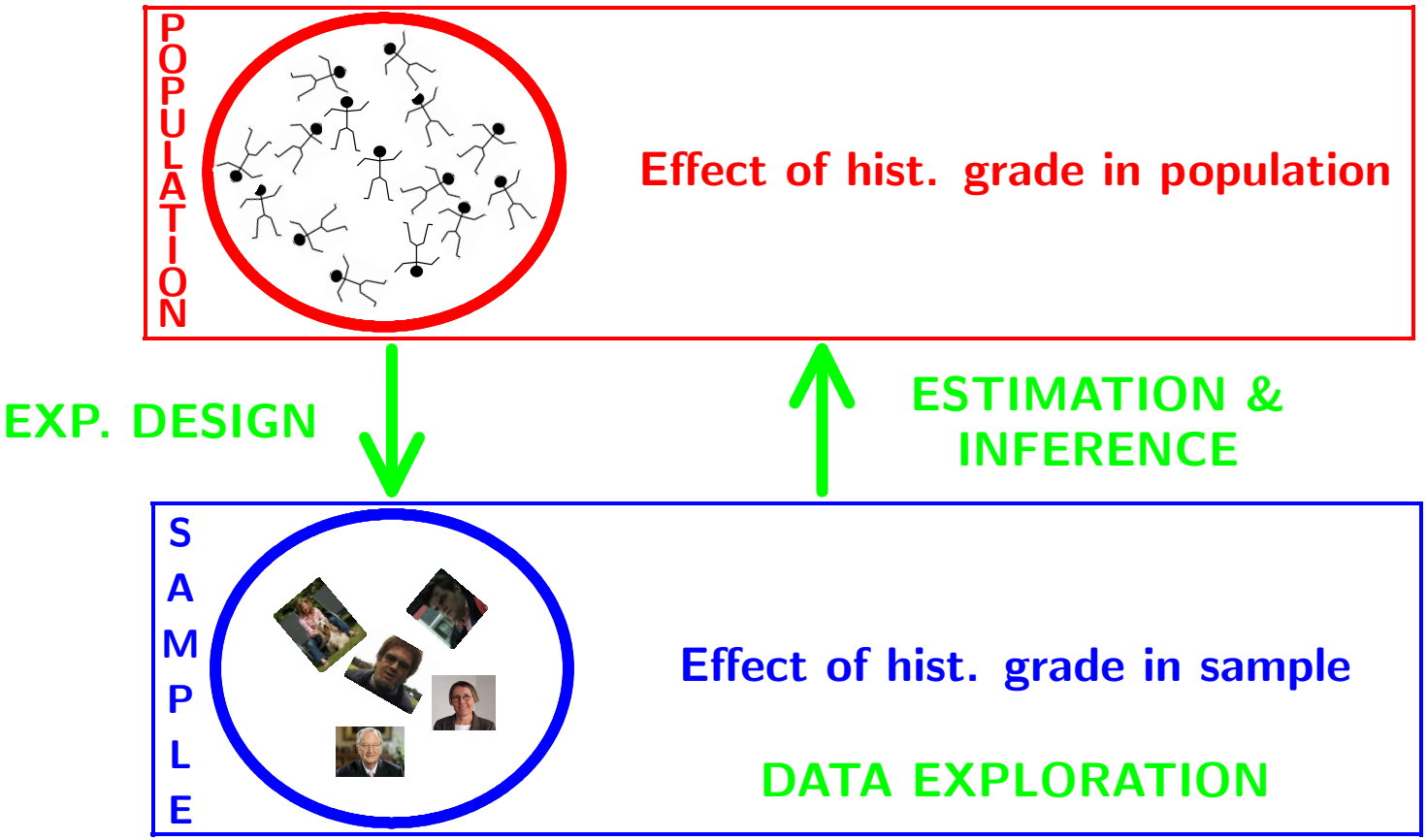
Breast cancer example (part of study https://doi.org/10.1093/jnci/djj052)

- Histologic grade in breast cancer clinically prognostic. Association of histologic grade on expression of KPNA2 gene that is known to be associated with poor BC prognosis.
- Population: all current and future breast cancer patients



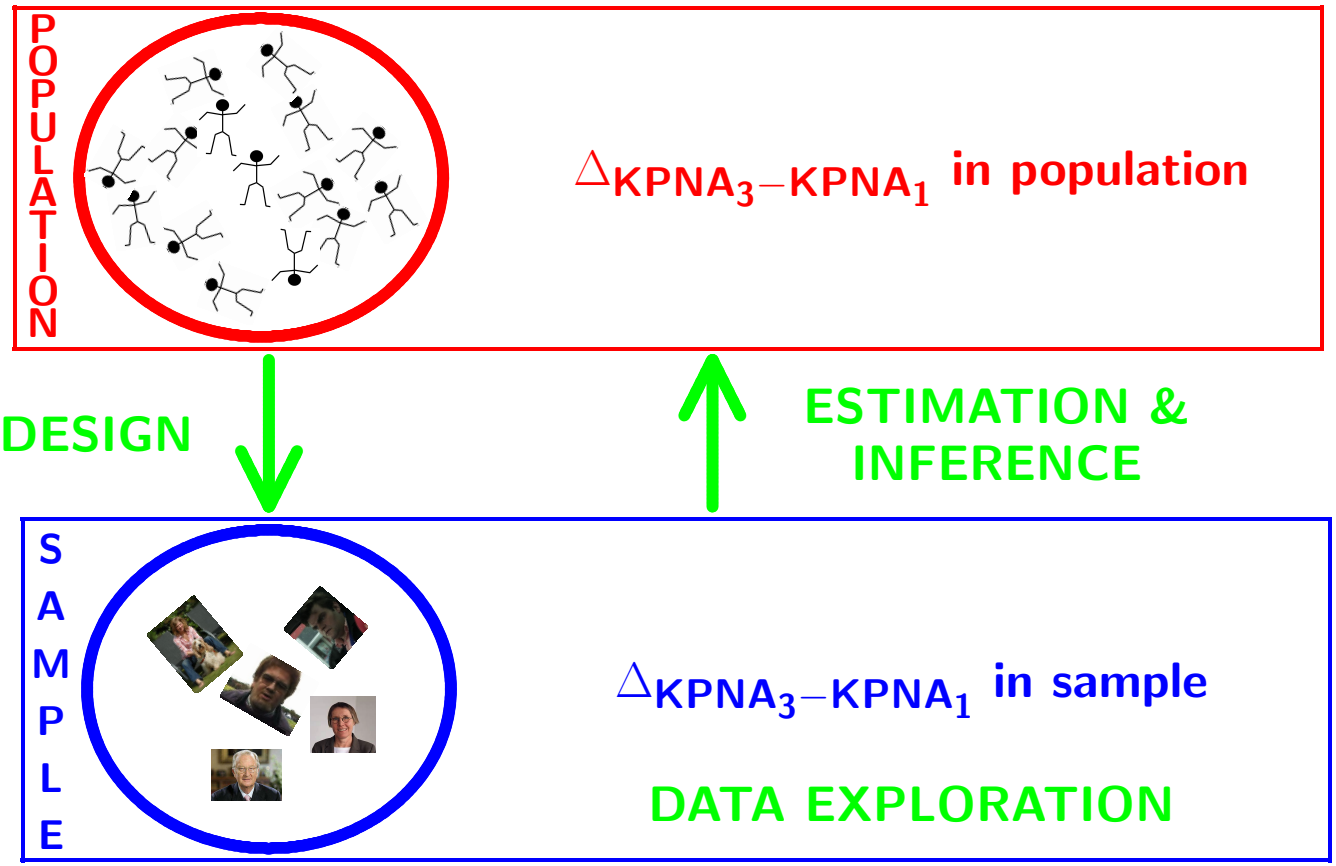
Breast cancer example (part of study https://doi.org/10.1093/jnci/djj052)

- Histologic grade in breast cancer clinically prognostic. Association of histologic grade on expression of KPNA2 gene that is known to be associated with poor BC prognosis.
- Population: all current and future breast cancer patients



Breast cancer example (part of study https://doi.org/10.1093/jnci/djj052)

- Histologic grade in breast cancer clinically prognostic. Association of histologic grade on expression of KPNA2 gene that is known to be associated with poor BC prognosis.
- Population: all current and future breast cancer patients



Data Exploration

Import gene data in R

```
> gene <- read.table("gse2990BreastcancerOneGene.txt",header=TRUE)
> head(gene)

   sample_name grade node size age     gene
28   OXFT_2221     3    1  5.5  76 367.8179
29    OXFT_209     3    1  2.5  66 590.3576
30   OXFT_1769     1    1  3.5  86 346.6583
31    OXFT_928     1    0  1.1  47 118.6996
32   OXFT_2093     1    1  2.2  74 519.4489
33   OXFT_1770     1    1  1.7  69 258.4455

> #transform the variable grade and node to a factor
> gene$grade <- as.factor(gene$grade)
> gene$node <- as.factor(gene$node)
```

Data Exploration

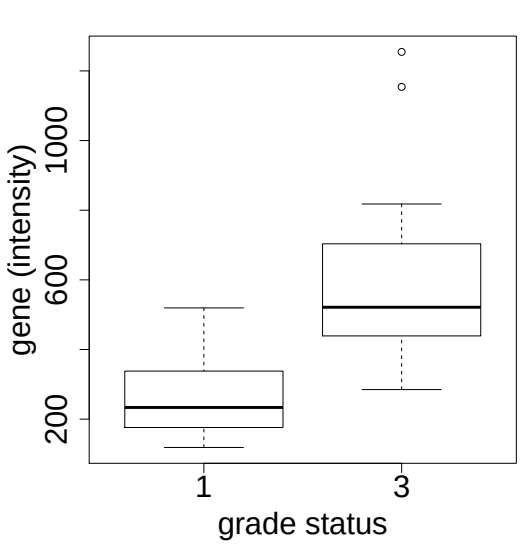
```
> gene$grade==1

 [1] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE
[13]  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
[25] FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE

> geneGrade1 <- subset(gene,grade==1)
> head(geneGrade1,3)

   sample_name grade node size age     gene
30   OXFT_1769     1    1  3.5  86 346.6583
31    OXFT_928     1    0  1.1  47 118.6996
32   OXFT_2093     1    1  2.2  74 519.4489

> geneGrade3 <- subset(gene,grade==3)
> head(geneGrade3,3)

   sample_name grade node size age     gene
28   OXFT_2221     3    1  5.5  76 367.8179
29    OXFT_209     3    1  2.5  66 590.3576
35   OXFT_1342     3    0  2.5  62 643.6799
```

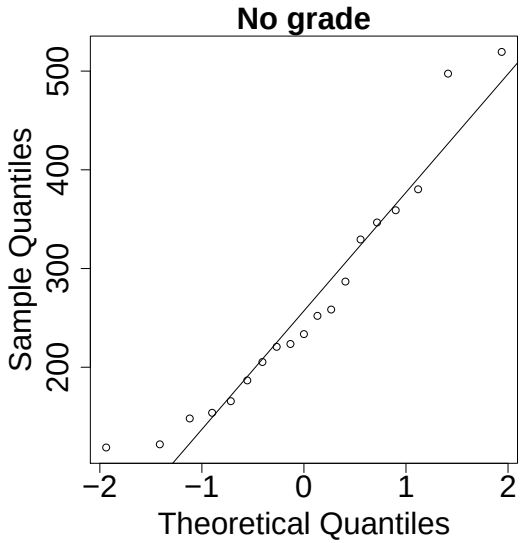
Data Exploration

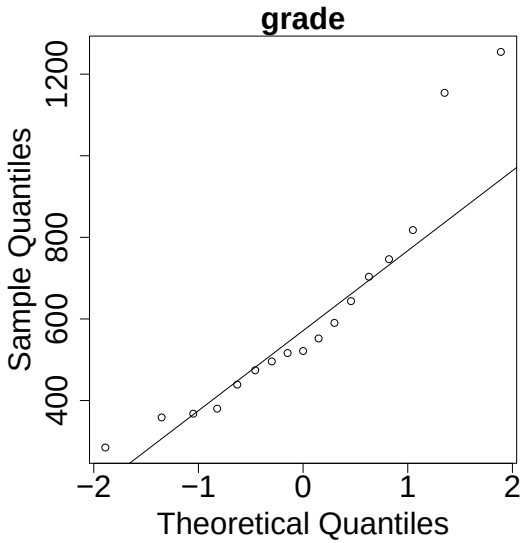
```
> mu1 <- mean(geneGrade1$gene)
> sd1 <- sd(geneGrade1$gene)
> se1 <- sd1/sqrt(nrow(geneGrade1))
> c(mu1,sd1,se1)

[1] 263.55165 116.55279  26.73904

> mu2 <- mean(geneGrade3$gene)
> sd2 <- sd(geneGrade3$gene)
> se2 <- sd2/sqrt(nrow(geneGrade3))
> c(mu2,sd2,se2)

[1] 605.96769 267.44027  64.86379
```

Data Exploration

```
> boxplot(gene~grade,data=gene,xlab="grade status", ylab="gene (intensity)",cex.main=2,cex.axis=2,cex.lab=
> qqnorm(geneGrade1$gene,main="No grade",cex.main=2,cex.axis=2,cex.lab=2)
> qqline(geneGrade1$gene,main="No grade")
> qqnorm(geneGrade3$gene,main="grade",cex.main=2,cex.axis=2,cex.lab=2)
> qqline(geneGrade3$gene,main="grade")
```

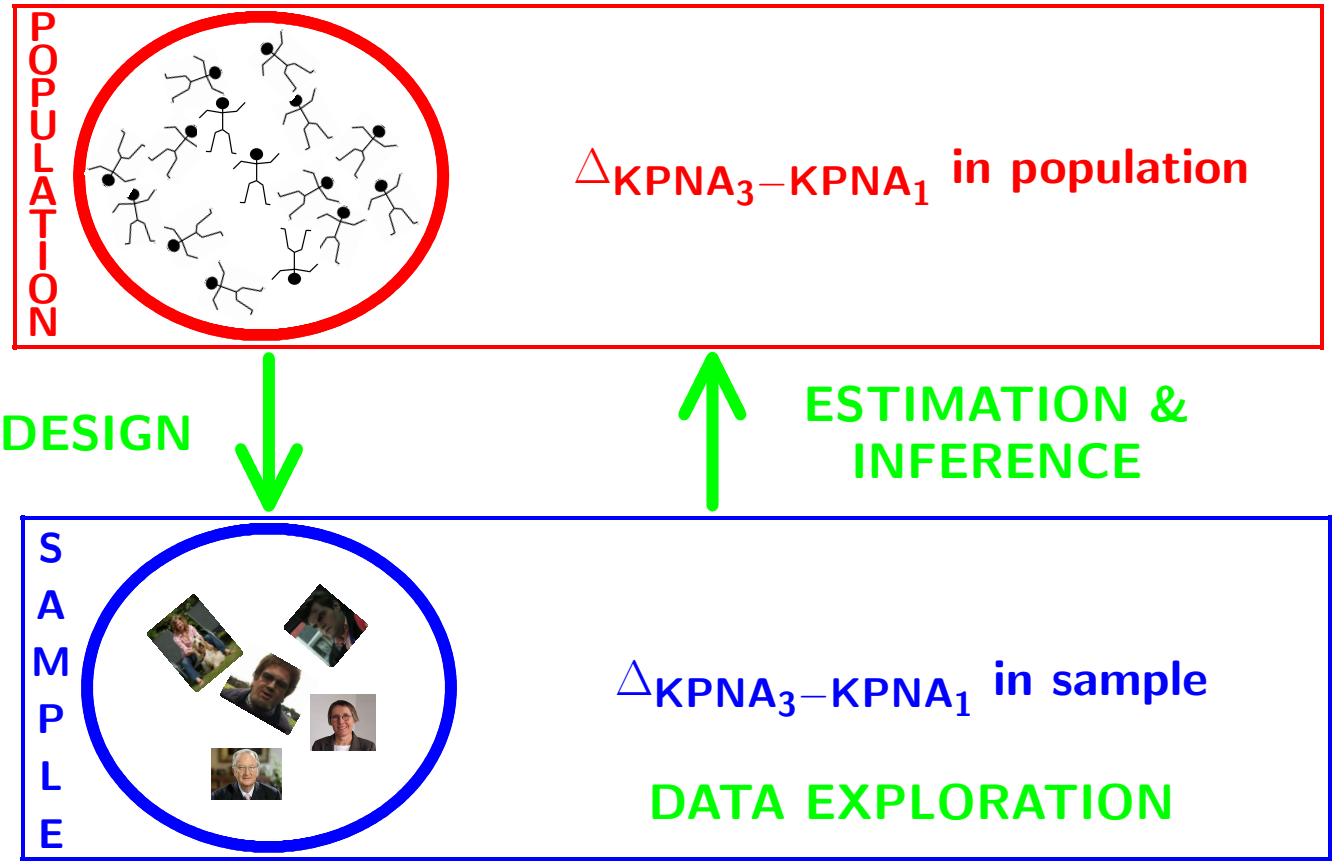






Breast cancer example

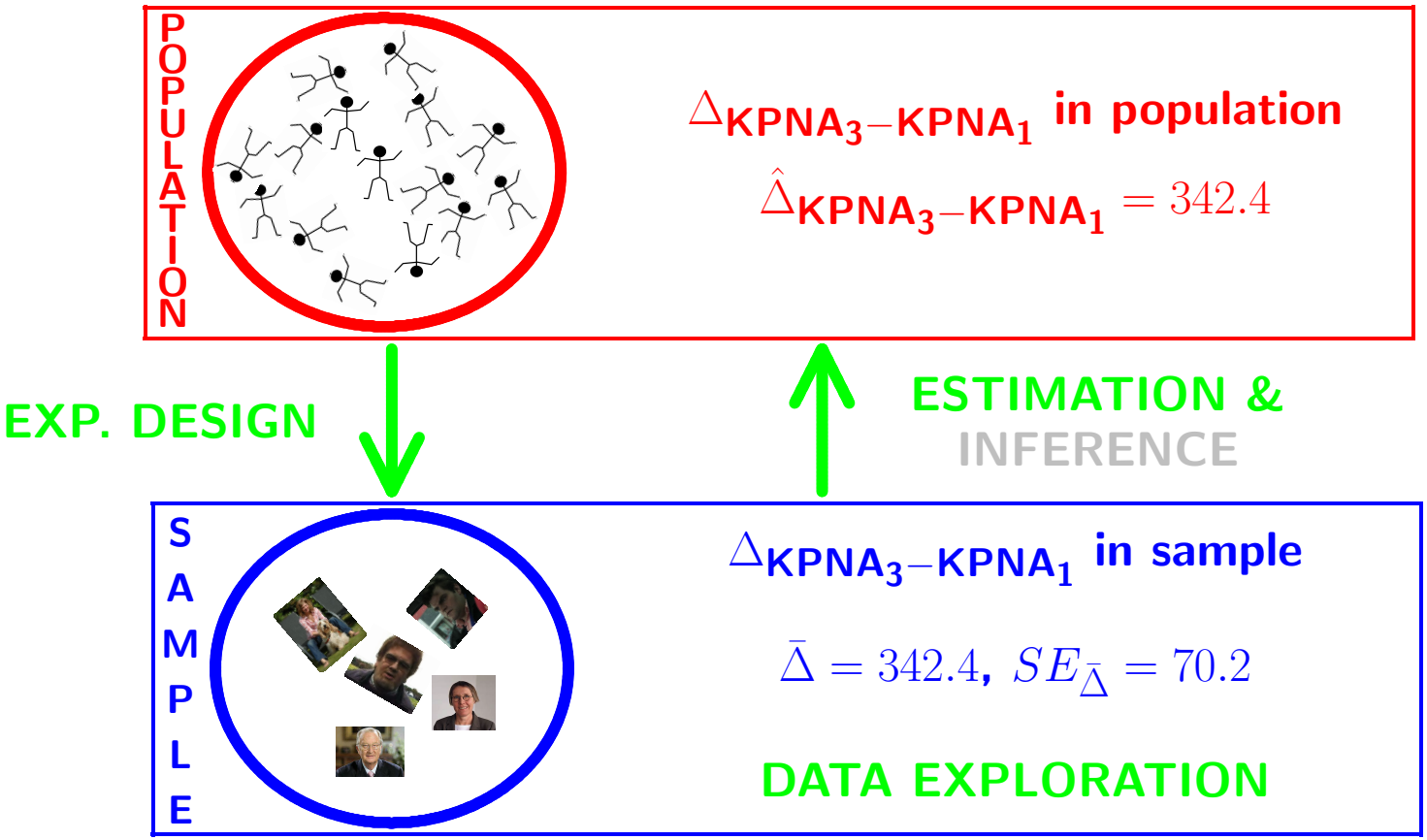
- Researchers want to assess the association of the histolical grade on KPNA2 gene expression



```
> Delta <- mu2-mu1
> seDelta <- sqrt(se1^2+se2^2)
> c(Delta,seDelta)

[1] 342.41604  70.15902
```

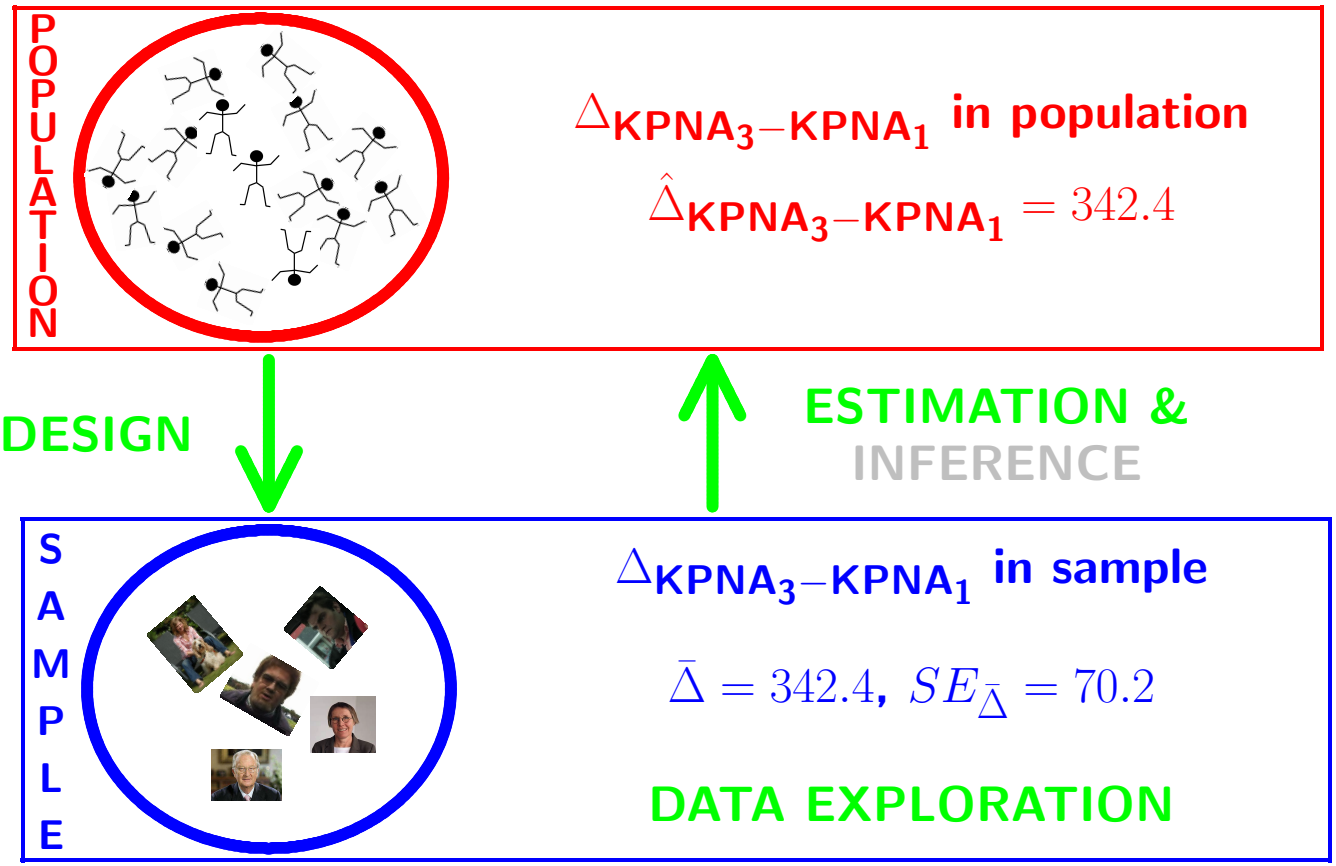
Breast cancer example

- Researchers want to assess the association of histological grade on KPNA2 gene expression



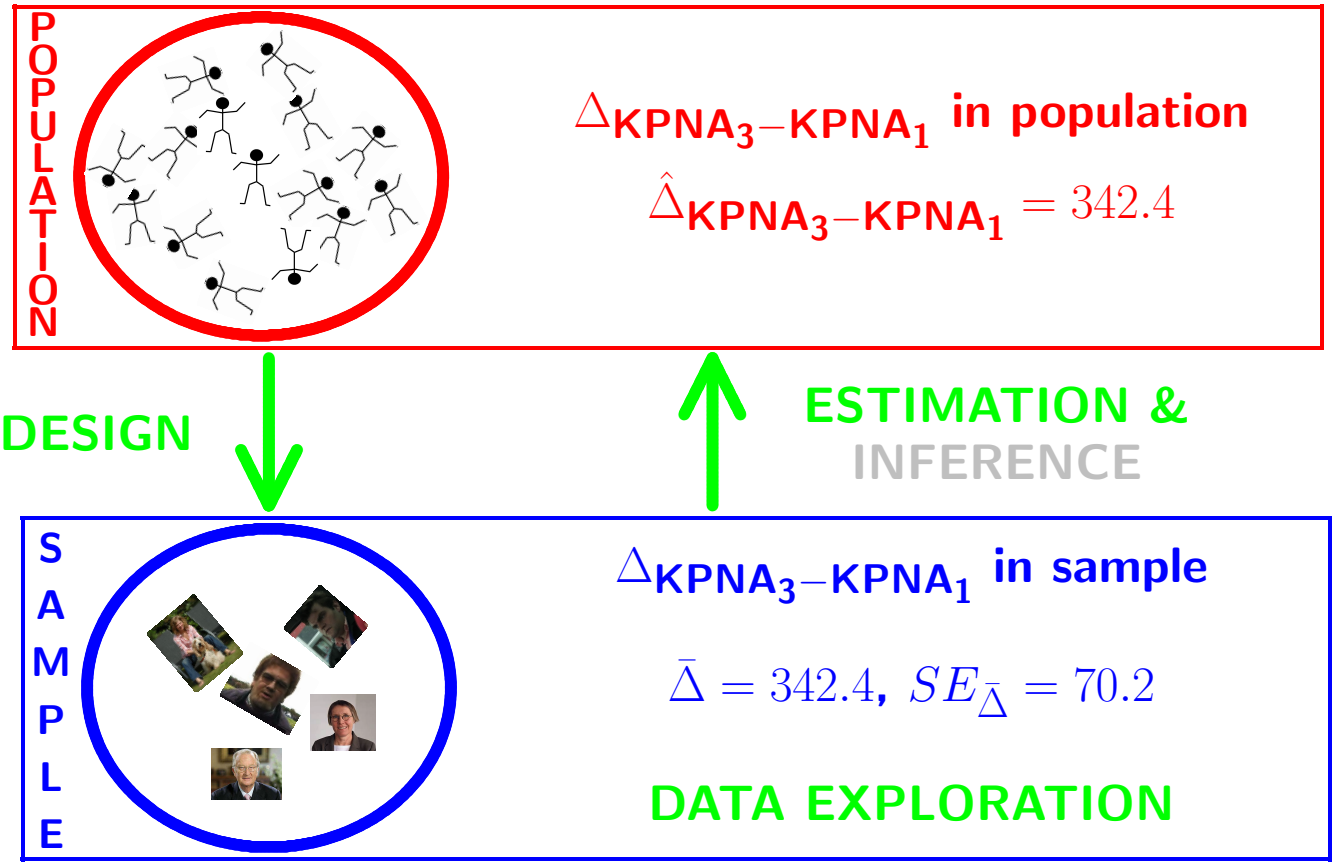
Breast cancer example

- Researchers want to assess the association of histological grade on KPNA2 gene expression
- Inference?
- testing + CI $\rightarrow$ Assumptions



Breast cancer example

- Researchers want to assess the association of grade on KPNA2 gene expression
- Inference?
- Statistical Test, which one?



Null hypothesis and alternative hypothesis

- In general we start from **alternative hypothese** H_A: we want to show an association
  - Gene expression of grade 1 and grade 3 patients is on average different

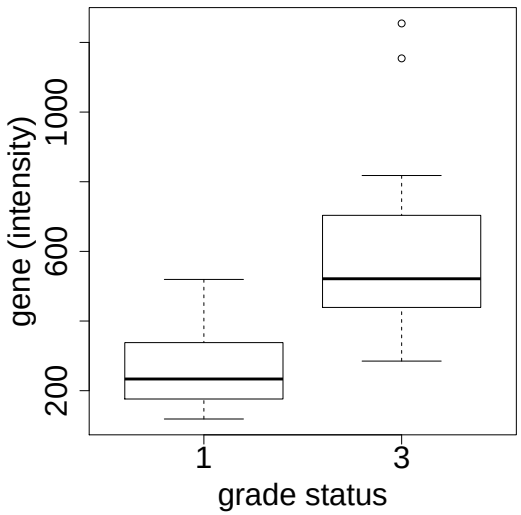
Null hypothesis and alternative hypothesis

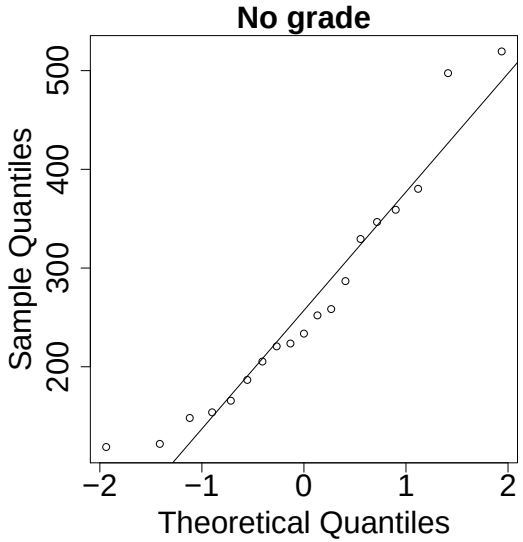
- In general we start from **alternative hypothese** H_A: we want to show an association
  - Gene expression of grade 1 and grade 3 patients is on average different
- But, we will assess it by falsifying the opposite: **null hypothesis** H_0
  - The average KPNA2 gene expression of grade 1 and grade 3 patients is equal

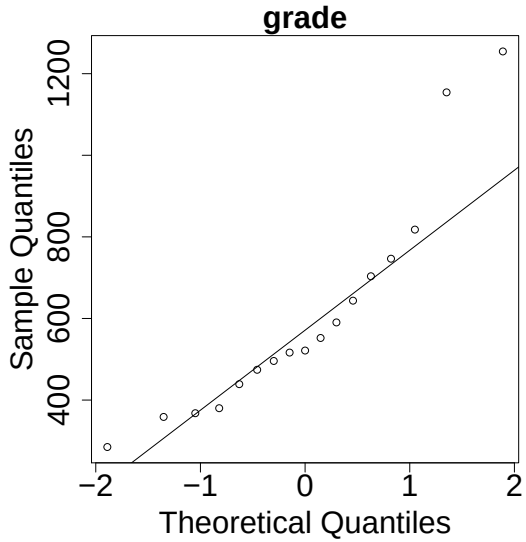
- How likely is it to observe an equal or more extreme association than the one observed in the sample when the null hypothesis is true?
- When we make assumptions about the distribution of our test statistic we can quantify this probability: **p-value**.
- If the p-value is below a significance threshold α we reject the null hypothesis
  We control the probability on a false positive result at the α-level (type I error)
- The p-value will only be calculated correctly if the underlying assumptions hold!

Analysis?

Analysis?



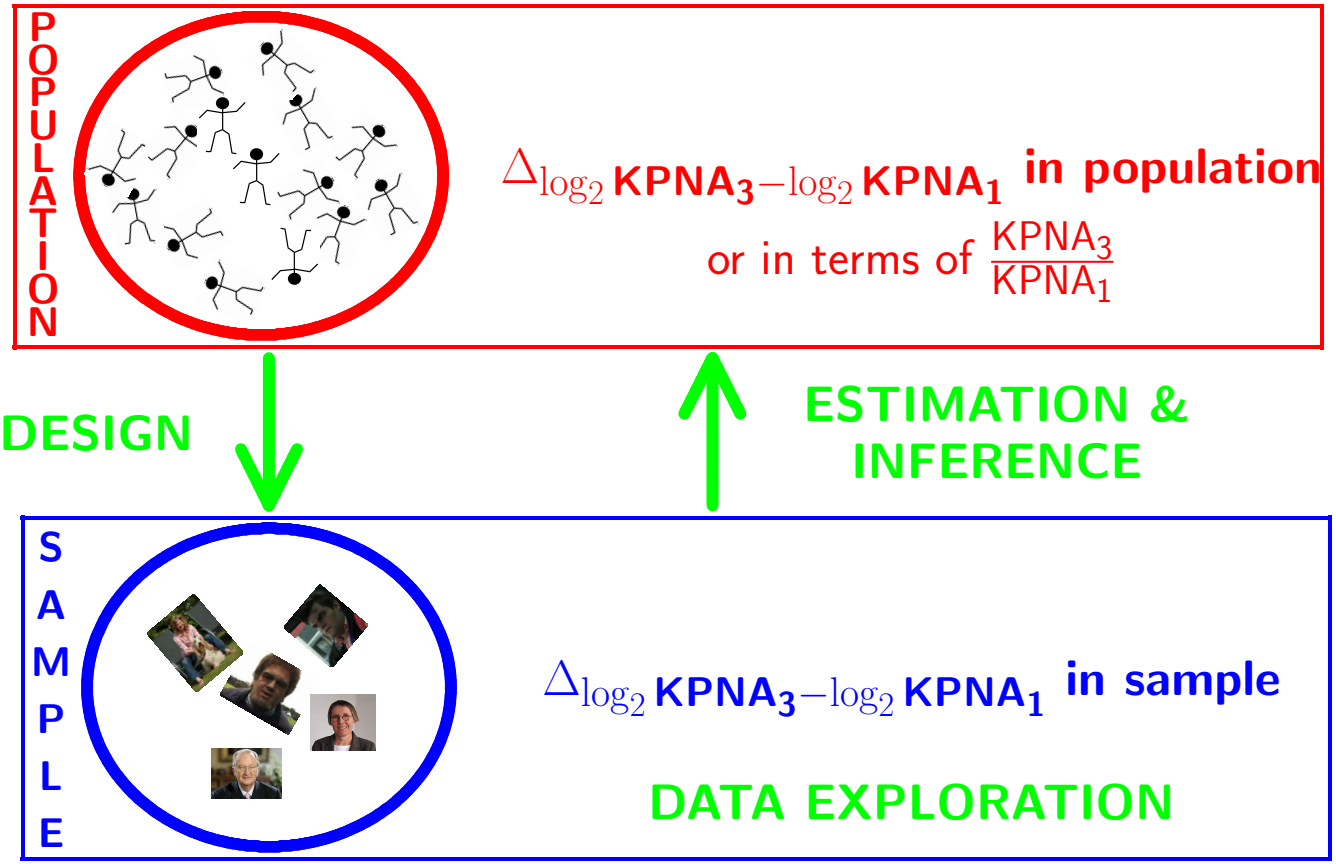





```
> t.test(gene~grade,data=gene)

^^IWelch Two Sample t-test

data:  gene by grade
t = -4.8806, df = 21.352, p-value = 7.61e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -488.1734 -196.6587
sample estimates:
mean in group 1 mean in group 3
       263.5516        605.9677

> tOwn <- (mu2-mu1)/sqrt(se1^2+se2^2)
> tOwn

[1] 4.88057

> pt(-abs(tOwn),21.352)*2

[1] 7.610148e-05
```

Breast cancer example

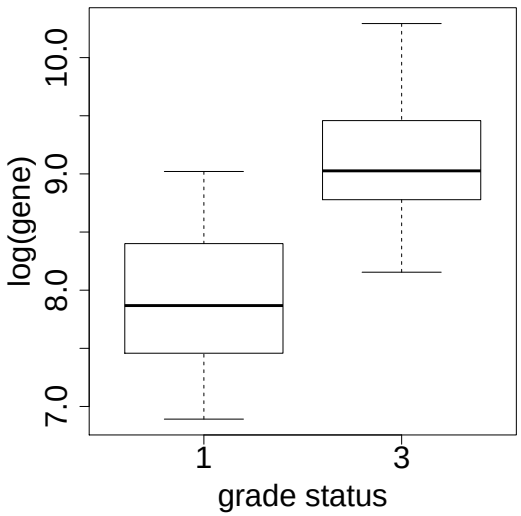
- Intensities are often not normally distributed and have a mean variance relation
- Commonly log-transformed

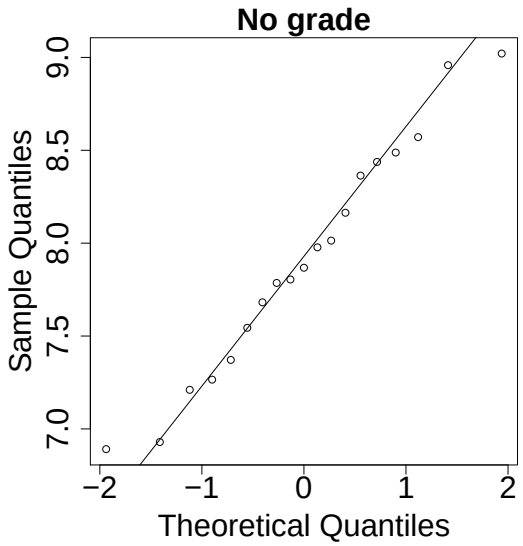


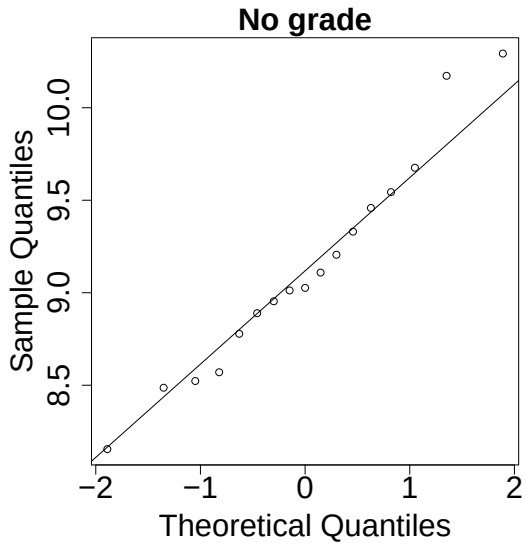
log-transformation

```
> gene$lgene <- log2(gene$gene)
> geneGrade1$lgene <- log2(geneGrade1$gene)
> geneGrade3$lgene <- log2(geneGrade3$gene)
> logtest <- t.test(lgene~grade,data=gene)
> logtest

^^IWelch Two Sample t-test

data:  lgene by grade
t = -6.0508, df = 33.962, p-value = 7.432e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6236927 -0.8072052
sample estimates:
mean in group 1 mean in group 3
       7.912963        9.128412

> logtest$estimate[2]-logtest$estimate[1]

mean in group 3
       1.215449

> 2^(logtest$estimate)

mean in group 1 mean in group 3
       241.0124        559.6621

> 2^(logtest$estimate[2]-logtest$estimat

mean in group 3
        2.32213

> 2^(logtest$conf.int)

[1] 0.3245038 0.5714879
attr(,"conf.level")
[1] 0.95

> 2^-(logtest$conf.int)

[1] 3.081628 1.749818
attr(,"conf.level")
[1] 0.95

> sqrt(var(geneGrade1$lgene)/nrow(geneGrade1)+var(geneGrade3$lgene)/nrow(geneGrade3))

[1] 0.200875
```

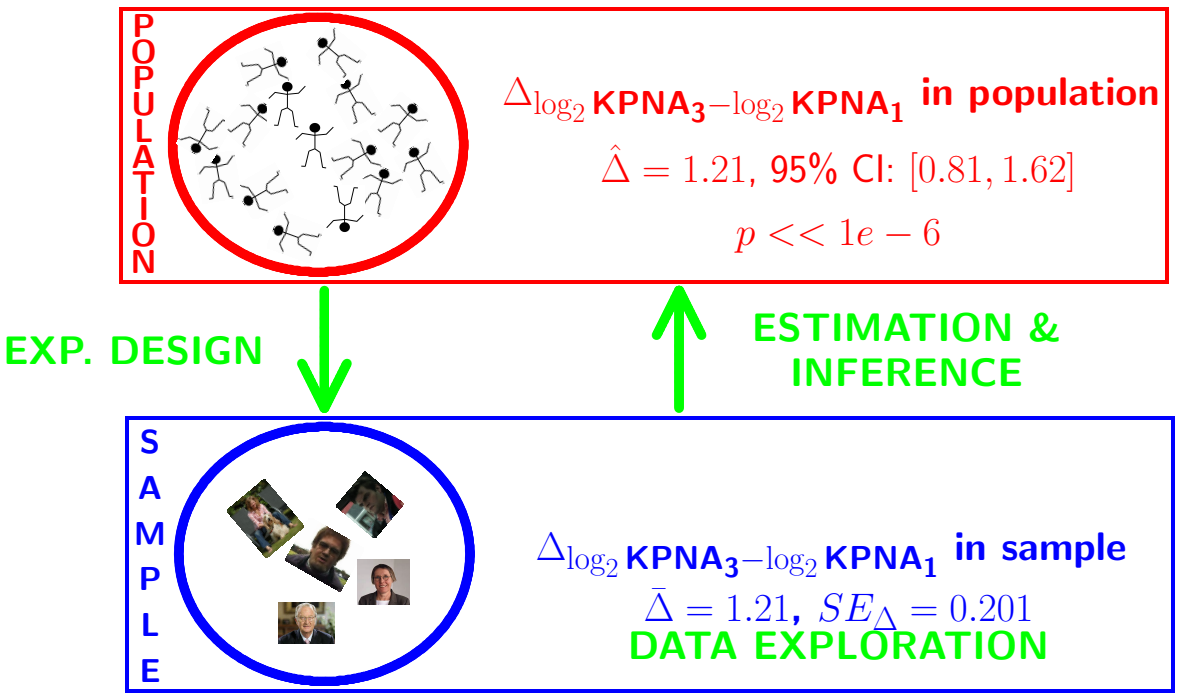
log-transformation



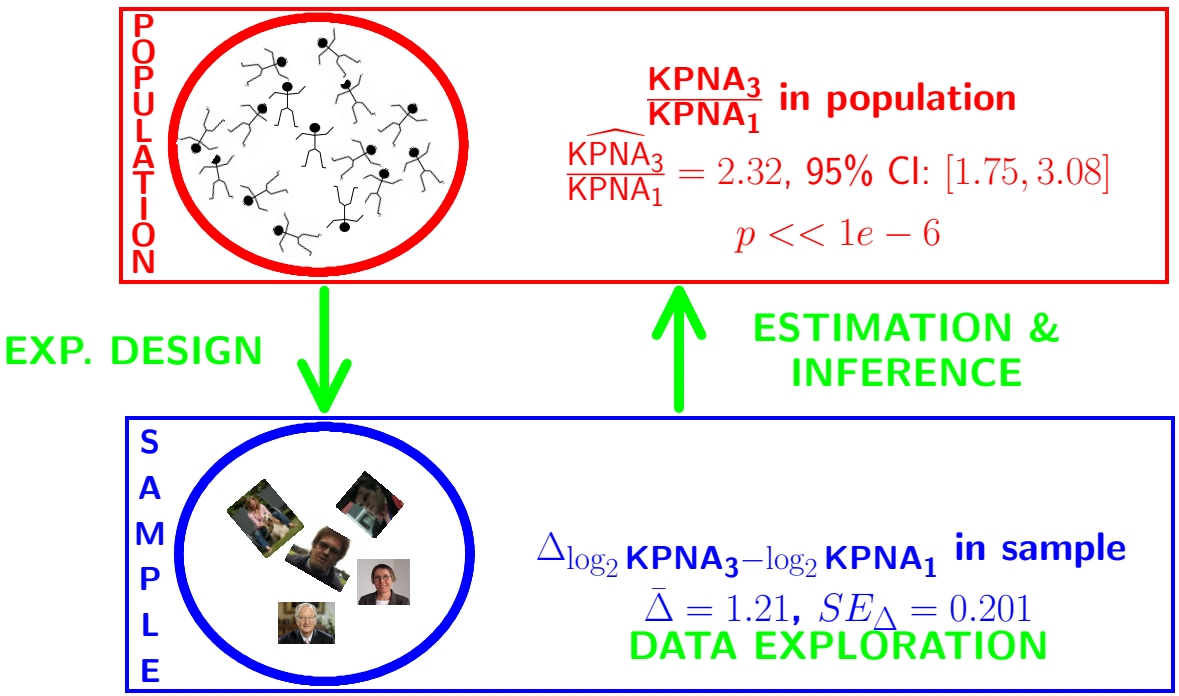




Breast cancer example

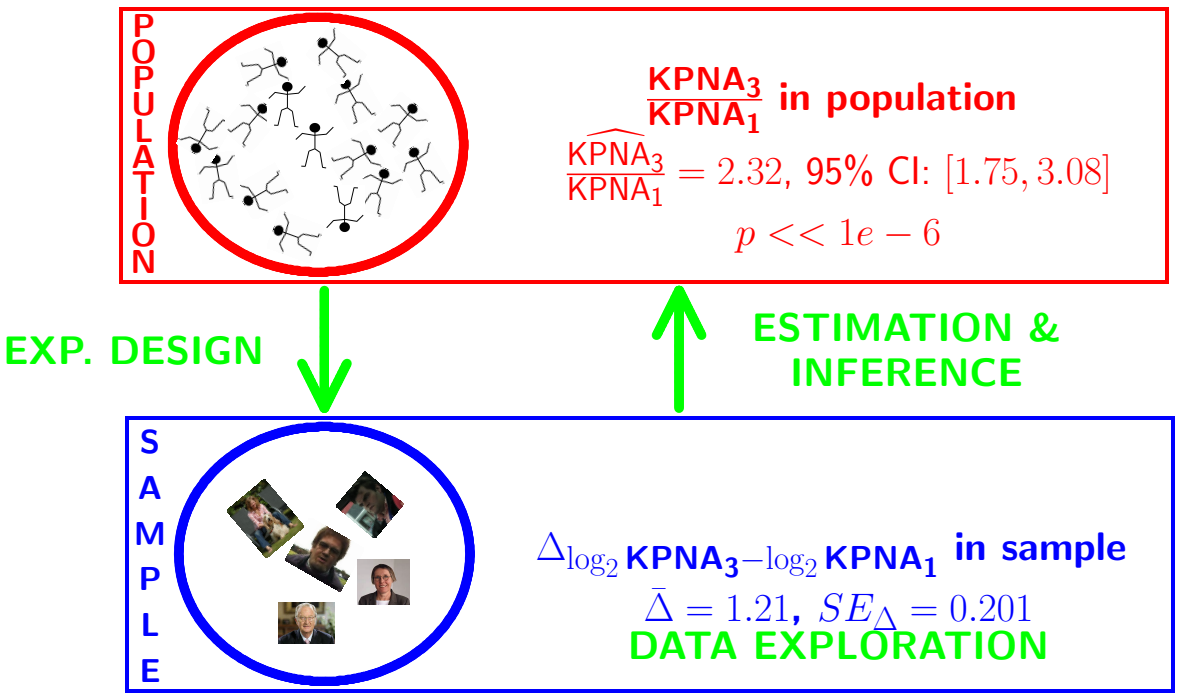


Breast cancer example



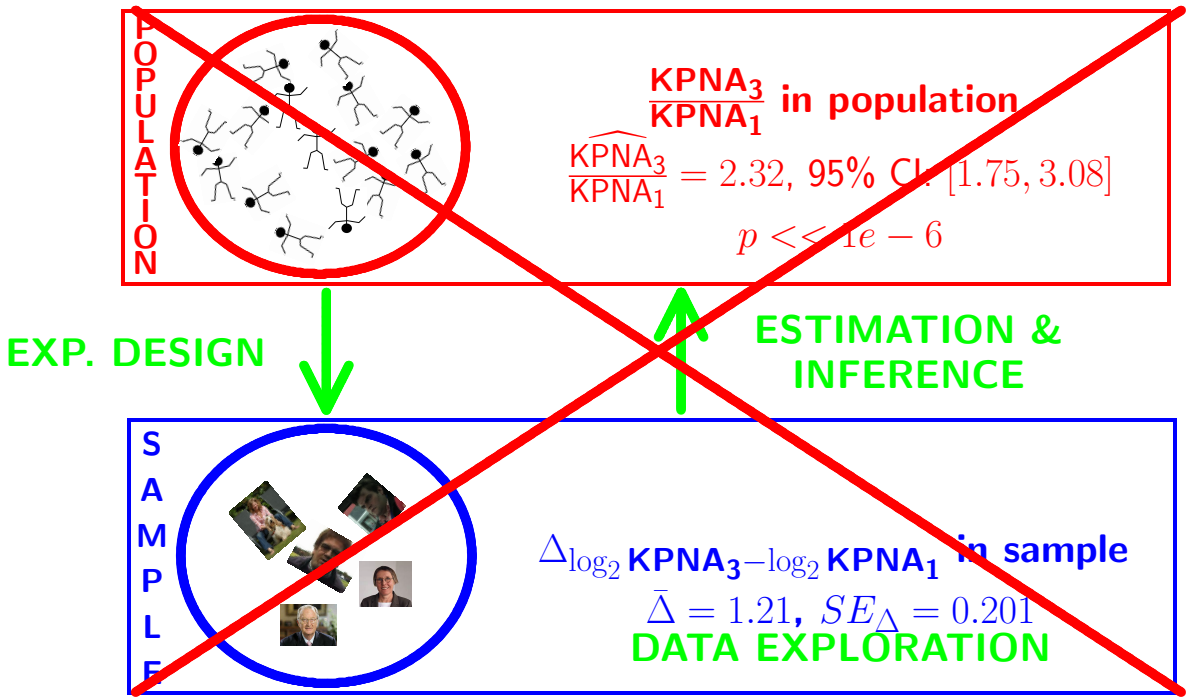
Breast cancer example

- There is a extremely significant association of the histological grade on the gene expression in tumor tissue. On average, the gene expression for the grade 3 patients is 2.32 times higher than the gene expression in grade 1 patients (95% CI [1.75,3.08], $p = 0.001$)



Breast cancer example

The patients also differ in the their lymph node status. Hence, we have a two factorial design: grade x lymph node status



SOLUTION?

General Linear Model

How can we integrate multiple factors and continuous covariates in linear model.

General Linear Model

How can we integrate multiple factors and continuous covariates in linear model.

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_{12} x_{i,1} x_{i,2} + \epsilon_i,$$

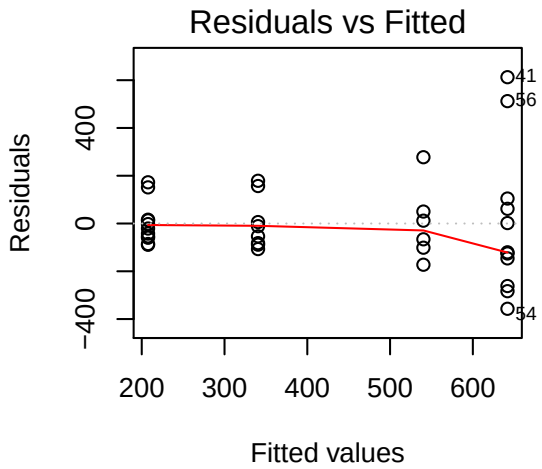
with

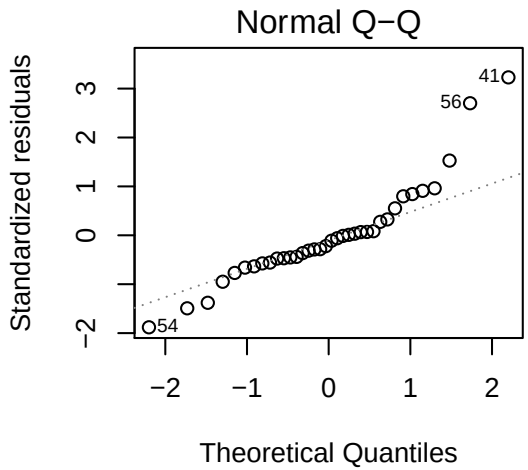
- $x_{i,1}$ a dummy variable for histological grade:
$$x_{i,1} = \begin{cases} 0 & \text{grade 1} \\ 1 & \text{grade 3} \end{cases}$$
- $x_{i,2}$ a dummy variable for :
$$x_{i,2} = \begin{cases} 0 & \text{lymph nodes were not removed} \\ 1 & \text{lymph nodes were removed} \end{cases}$$
- ϵ_i?

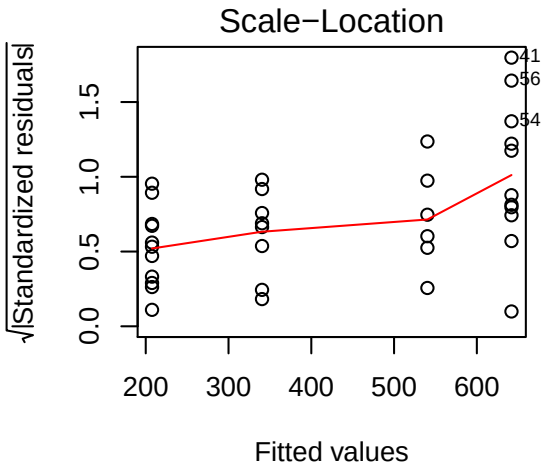
General Linear Model

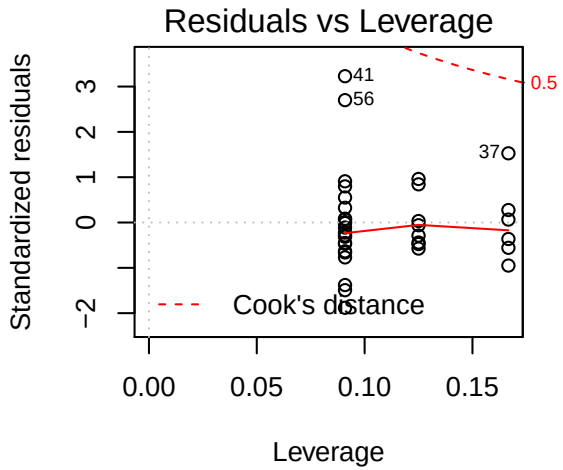
```
> lm1 <- lm(gene~grade*node,data=gene)
> summary(lm1)

Call:
lm(formula = gene ~ grade * node, data = gene)

Residuals:
    Min      1Q  Median      3Q     Max
-356.85  -91.98  -31.47   53.00  612.73

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    207.60      60.00   3.460  0.00155 **
grade3         434.21      84.85   5.117 1.41e-05 ***
node1          132.88      92.46   1.437  0.16040
grade3:node1  -234.43     136.92  -1.712  0.09655 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 199 on 32 degrees of freedom
Multiple R-squared:  0.4809,^^IAdjusted R-squared:  0.4322
F-statistic: 9.881 on 3 and 32 DF,  p-value: 9.181e-05
```

General Linear Model (problems?)









Breast cancer example

- Paper: https://doi.org/10.1093/jnci/djj052
- Histologic grade in breast cancer provides clinically important prognostic information. Two factors have to be concidered: Histologic grade (grade 1 and grade 3) and lymph node status (0 vs 1). The researchers assessed gene expression of the KPNA2 gene a protein-coding gene associated with breast cancer and are mainly interested in the association of histological grade. Note, that the gene variable consists of background corrected normalized intensities obtained with a microarray platform. Upon log-transformation, they are known to be a good proxy for the log transformed concentration of gene expression product of the KPNA2 gene.
- Research questions and translate them towards model parameters (contrasts)?
- Make an R markdown file to answer the research questions

Linear regression in matrix form

Linear Regression (LR)

- Consider a vector of predictors $\mathbf{x} = (x_1, \ldots, x_p)$ and
- a real-valued response Y
- then the linear regression model can be written as

$$Y = f(\mathbf{x}) + \epsilon = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \epsilon$$

with i.i.d. $\epsilon \sim N(0, \sigma^2)$

- n observations $(\mathbf{x}_1, y_1) \dots (\mathbf{x}_n, y_n)$
- Regression in matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\text{with } \mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\text{and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Least Squares (LS)

- Minimize the residual sum of squares

$$\begin{aligned} RSS(\boldsymbol{\beta}) &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \end{aligned}$$

- or in matrix notation

$$\begin{aligned} RSS(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \end{aligned}$$

with the L_2-norm of a p-dim. vector v $\|\mathbf{v}\| = \sqrt{v_1^2 + \ldots + v_p^2}$

$\rightarrow$ $\hat{\boldsymbol{\beta}} = \text{argmin}_\beta \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$

Minimize RSS

$$\begin{aligned}
\frac{\partial RSS}{\partial \boldsymbol{\beta}} &= \mathbf{0} \\
\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \mathbf{0} \\
-2\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) &= \mathbf{0} \\
\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^T\mathbf{Y} \\
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}
\end{aligned}$$

Variance Estimator?

$$
\begin{aligned}
\hat{\Sigma}_{\hat{\beta}} &= \text{var}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\right] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{var}\left[\mathbf{Y}\right]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}\sigma^2)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{I}\quad\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2
\end{aligned}
$$

Homework: Adopt the gene analysis in matrix form. Calculate

- model parameters and contrasts of interest
- standard errors, standard errors on contrasts
- t-test statistics on the model parameters and contrasts of interest
- compare your results with the output of the lm(.) function
- details on the implementation can be found in the book of Faraway (chapter 2).

Design Matrix:

```
> X <- model.matrix(~grade*node,data=gene)
> head(X)

   (Intercept) grade3 node1 grade3:node1
28           1      1     1            1
29           1      1     1            1
30           1      0     1            0
31           1      0     0            0
32           1      0     1            0
33           1      0     1            0
```

- Transpose of a matrix: use function t(.)

```
> t(X)
             28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
(Intercept)   1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
grade3        1  1  0  0  0  0  0  1  0  1  0  1  0  1  1  0  1  0  0  0  0  1
node1         1  1  1  0  1  1  0  0  0  1  1  0  0  0  0  0  0  0  0  0  0  1
grade3:node1  1  1  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  1
             50 51 52 53 54 55 56 57 58 59 61 62 63 64
(Intercept)   1  1  1  1  1  1  1  1  1  1  1  1  1  1
grade3        0  1  1  0  1  1  1  1  0  1  0  0  1  0
node1         1  0  0  1  0  1  0  0  1  0  0  1  1  0
grade3:node1  0  0  0  0  0  1  0  0  0  0  0  0  1  0
attr(,"assign")
[1] 0 1 2 3
attr(,"contrasts")
attr(,"contrasts")$grade
[1] "contr.treatment"

attr(,"contrasts")$node
[1] "contr.treatment"
```

- Invert matrix: use function solve(.)
- Diagonal elements of a matrix: use function diag(.)
- Matrix product % * % operateor

```
> c(lm1$fitted)[1:5]
      28       29       30       31       32
540.2553 540.2553 340.4795 207.6041 340.4795
> t(X%*%lm1$coef)[1:5]
[1] 540.2553 540.2553 340.4795 207.6041 340.4795
```

```
> summary(lm1)

Call:
lm(formula = gene ~ grade * node, data = gene)

Residuals:
    Min      1Q  Median      3Q     Max
-356.85  -91.98  -31.47   53.00  612.73

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    207.60      60.00   3.460  0.00155 **
grade3         434.21      84.85   5.117 1.41e-05 ***
node1          132.88      92.46   1.437  0.16040
grade3:node1  -234.43     136.92  -1.712  0.09655 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 199 on 32 degrees of freedom
Multiple R-squared:  0.4809,^^IAdjusted R-squared:  0.4322
F-statistic: 9.881 on 3 and 32 DF,  p-value: 9.181e-05

> dfRes <- (nrow(X)-ncol(X))
> varRes <- sum(((gene$gene)-X%*%lm1$coef)^2)/dfRes
> c(dfRes,sqrt(varRes))

[1]  32.0000 198.9893
```

```
> summary(lm1)$cov.unscaled

             (Intercept)      grade3       node1 grade3:node1
(Intercept)   0.09090909 -0.09090909 -0.09090909   0.09090909
grade3       -0.09090909  0.18181818  0.09090909  -0.18181818
node1        -0.09090909  0.09090909  0.21590909  -0.21590909
grade3:node1  0.09090909 -0.18181818 -0.21590909   0.47348485

> solve(t(X)%*%X)

             (Intercept)      grade3       node1 grade3:node1
(Intercept)   0.09090909 -0.09090909 -0.09090909   0.09090909
grade3       -0.09090909  0.18181818  0.09090909  -0.18181818
node1        -0.09090909  0.09090909  0.21590909  -0.21590909
grade3:node1  0.09090909 -0.18181818 -0.21590909   0.47348485
```

Extract diagonal elements from matrix

```
> diag(solve(t(X)%*%X))

 (Intercept)       grade3        node1 grade3:node1
  0.09090909   0.18181818   0.21590909   0.47348485
```