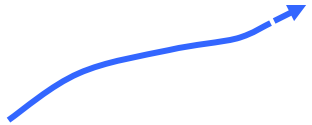
Predictive modeling for cancer prognosis

Gene expression data

mRNA Samples

Gene	sample1	sample2	sample3	sample4	sample5	...
1	0.46	0.30	0.80	1.51	0.90	...
2	-0.10	0.49	0.24	0.06	0.46	...
3	0.15	0.74	0.04	0.10	0.20	...
4	-0.45	-1.03	-0.79	-0.56	-0.32	...
5	-0.06	1.06	1.35	1.09	-1.09	...

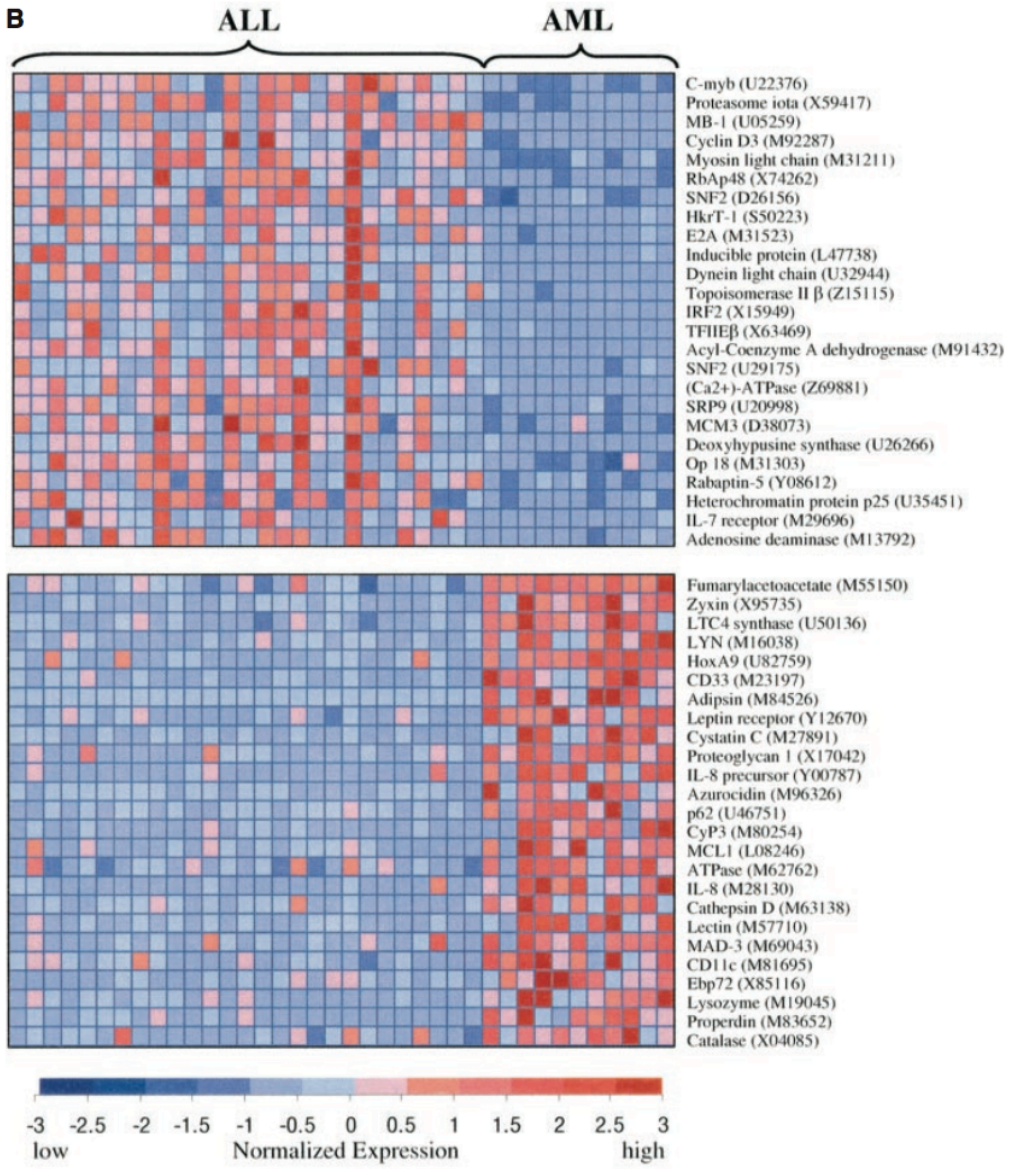


gene-expression level or ratio for gene *i* in mRNA sample *j*

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Golub, et al., Science 286:531-537 (1999).



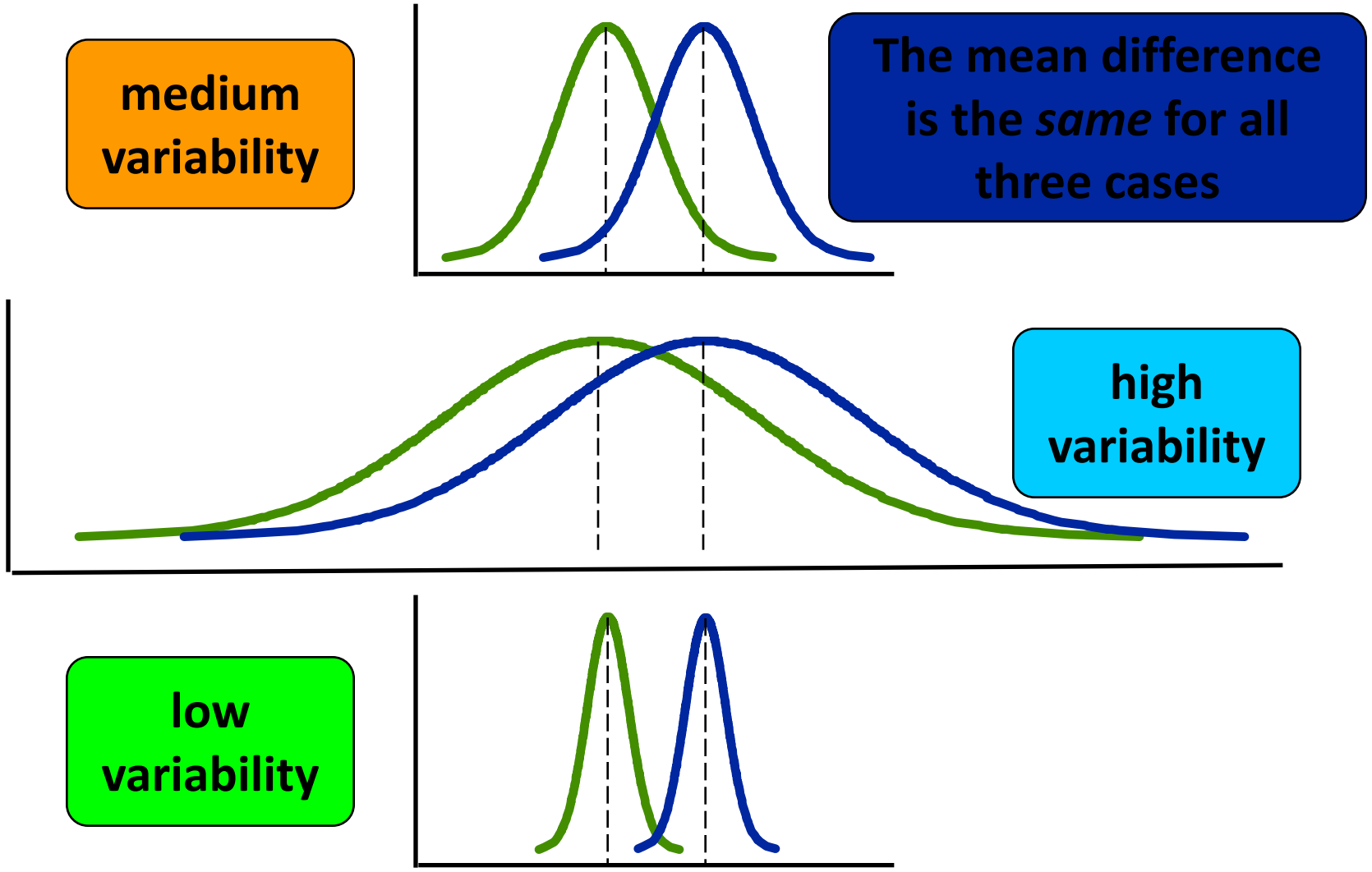
Differential (supervised) expression analysis

- Two (or more) classes of sample
  - Unpaired or paired
  - Perhaps defined from a clustering analysis
- Continuous variable
- Survival outcome (censored data)
- What genes are significantly different between classes
  - Or correlated with a variable
  - Or associated with outcome

Basic Data Analysis

- Differential expression analysis (what differs between 2 or more sample types)
  - Fold change (relative increase or decrease in intensity for each gene)
  - T-test type statistics
  - P-values, multiple hypothesis testing, false discovery rates
- Clustering samples and/or genes by similar patterns across the dataset

What does *difference* mean?



The “p-value”

- p-value = P(Type I Error) = P(Reject H_0 | H_0 is true)
- This is also called the “statistical significance.”
- **It represents an acceptable probability of making a mistake (i.e. false positive)!**

- A p-value of 0.05 or less customarily treated as an “acceptable” error level (this is rather arbitrary)

- Also important to consider type II error (false negative)

- Don’t be a p-value slave or p-hacker!
  - https://med.stanford.edu/news/all-news/2016/03/misleading-p-values-showing-up-more-often-in-journals.html
  - http://www.nature.com/articles/nmeth.4120

Type I and Type II Errors for Single Hypothesis Test

Decision	Actual Situation "Truth": H_0 True	Actual Situation "Truth": H_0 False
Do Not Reject H_0	Correct Decision $1 - \alpha$	Incorrect Decision Type II Error β
Rejct H_0	Incorrect Decision Type I Error α	Correct Decision $1 - \beta$

$$\alpha = P(Type\ I\ Error) \quad \beta = P(Type\ II\ Error)$$

H0 is the null hypothesis e.g. that there is no difference between two groups

Multiple hypothesis testing

- Testing 1000’s of genes at once
- Normal p-value not sufficient
  - Many results will be false positives

- Modifications that account for this:
  - False discovery rate
    - Number of positive results that are likely to be wrong

Why Multiple Testing Matters

- Genomics = Lots of Data = Lots of Hypothesis Tests
- A typical gene expression expression might result in performing 20,000 separate hypothesis tests.
- If we use a standard p-value cut-off of 0.05, we would expect **1000** genes to be deemed "significant" by chance.

Identifying Differential Expression

- Compare treatment to the control
  - The fold approach
  - The t-test (unpaired or paired)
  - Variations of the t-test
    - SAM: significance analysis of microarrays
- Compare several treatments
  - ANOVA: analysis of variance
  - MAANOVA: http://www.jax.org/staff/churchill/labsite/software/anova/index.html

Fold Change

- Measure ratios of gene expression levels.
- Ratio = T_i/C_i. Ratio of measured treatment intensity to control intensity for the i^{th} gene
- The $\log_2$ ratio treats up and down regulated genes equally
  - e.g. when looking for genes with more than 2 fold variation in expression

The Fold Approach

- In northern blot analysis, a 2-fold change can be seen with bare eyes
- Thus biologists tend to use 2-fold as the threshold of differential expression

- If x1 and x2 are in log2 space:
  - Difference in mean(x_1, x_2) > 1
  - Difference in mean(x_1, x_2) < -1

Two-fold up-regulation

- Problems with this approach:
  - Only identifies most changed genes.
  - Also identifies noise and highly variable genes.
  - Ratio is unstable when the denominator is small.

  - No estimate of significance of the results

Ratios are unstable

- Initial measurements:

  30/60 = 0.5

  500/1000 = 0.5

- Add random noise (+15 numerator and -15 denominator):

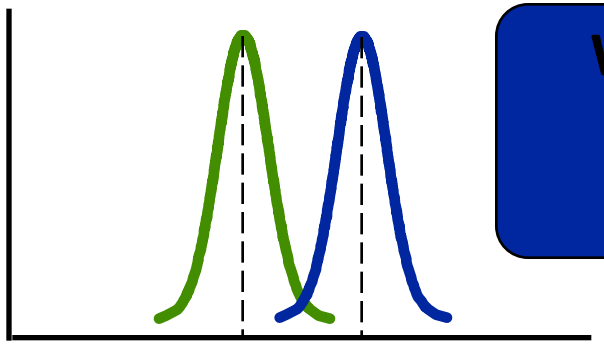
  45/45 = 1.0

  515/985 = 0.52

What does *difference* mean?

- a statistical difference is a function of the *difference between means* relative to the *variability*
- a small difference between means with large variability could be due to *chance*
- like a *signal-to-noise* ratio

low variability



Which one shows the *greatest* difference?

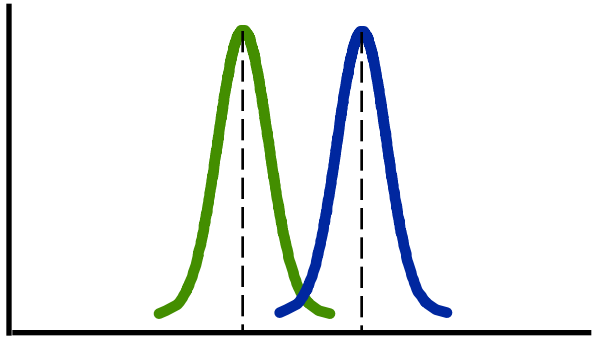
So we estimate

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

$$= \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)}$$

$$= \text{t-value}$$

low variability



The Student's t-test

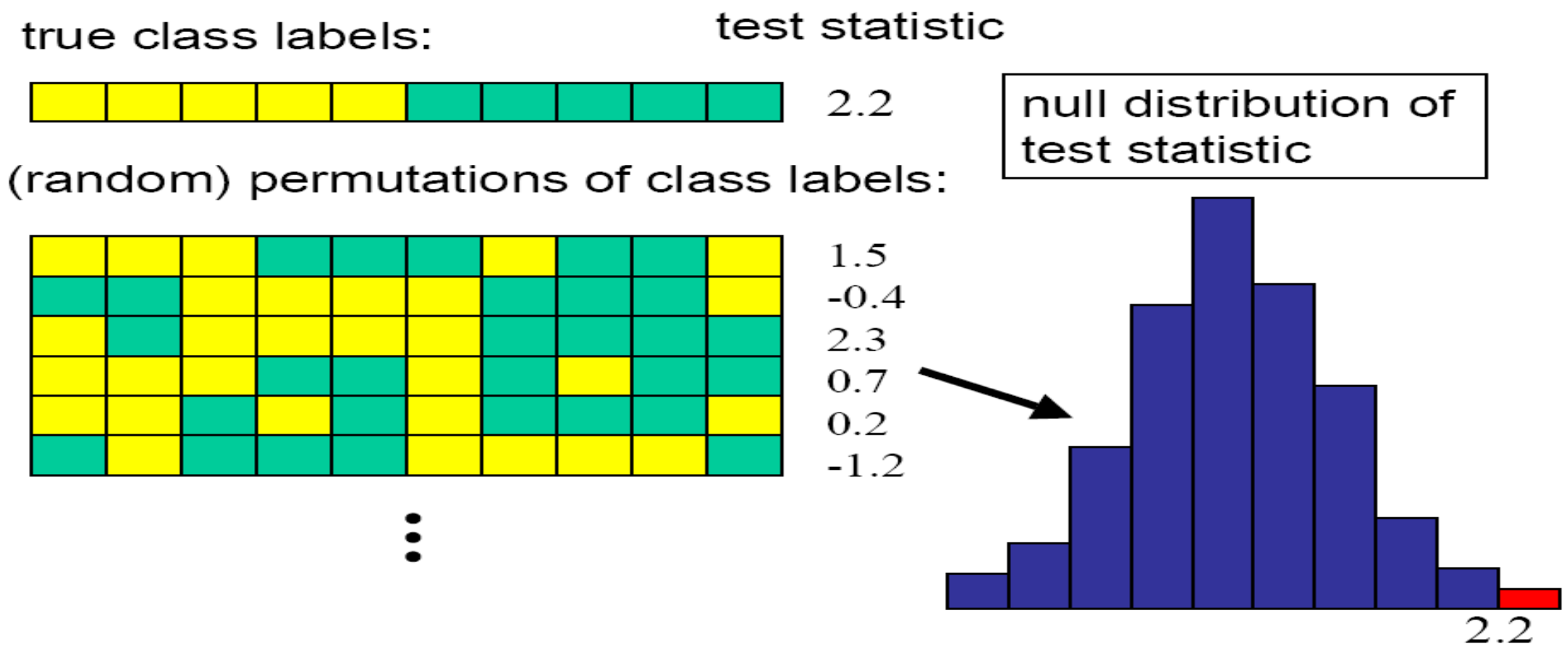
- We make use of the T-distribution distribution in the two-sample Students t-test.
- This test is used to test whether two samples come from distributions with the same means.
- The samples are assumed to come from Gaussian (normal) distributions.
- The two samples must have similar dispersions

Student's t-distribution

- **is mound shaped**
- **is symmetrical about zero**
- **is more widely dispersed than the standard normal distribution**
- **it's actual shape is dependent on the sample size**

- **different t distributions are identified by their degrees of freedom (df), where df = n-1**

Permutation test



Significance analysis of microarrays

- “Significance analysis of microarrays applied to the ionizing radiation response” Tusher, Tibshirani, Chu PNAS 2001
- T-test combined with permutation testing to generate false discovery rate

- Convenient Excel and R packages:
  - https://statweb.stanford.edu/~tibs/software.html
  - (or through Bioconductor)

Significance analysis of microarrays applied to the ionizing radiation response

- Lymphoblastoid cell lines grown with (n=4) or without irradiation
  - Small sample size! Generally need more
- 6800 genes
  - Old technology
- Test statistic for each gene

$$d(i) = \frac{\bar{x}_{\mathrm{I}}(i) - \bar{x}_{\mathrm{U}}(i)}{s(i) + s_0}$$

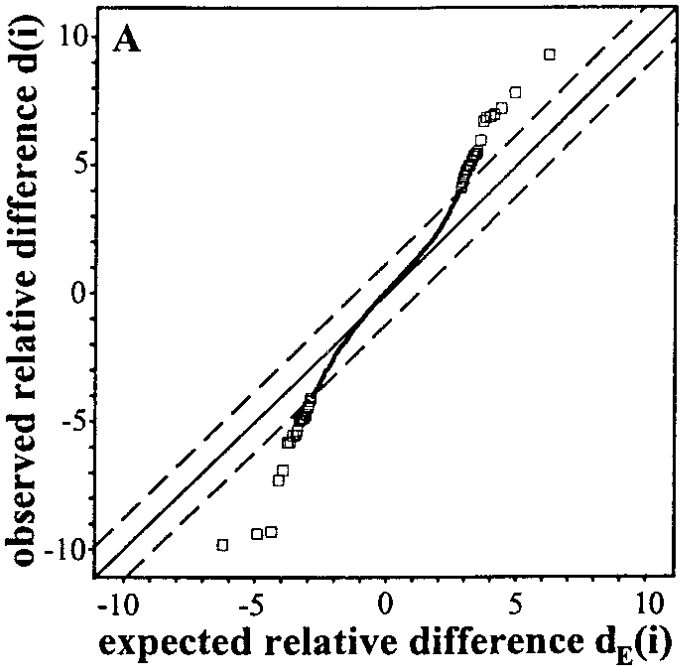
$$s(i) = \sqrt{a\left\{\sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2\right\}}$$

SAM procedure

- Rank genes by d(i)
- Do the same after permuting sample labels -> dE(i)
  - In this case there are only 36 balanced permutations
  - Generally want thousands (need more than you think)

- Identify genes for which d(i) is greater than expected by some amount Δ

- At a specified threshold Δ, the false discovery rate is # of genes in the permutations that appear significant
- Choose Δ according to the FDR you want
- Omitting s0 -> higher FDRs…

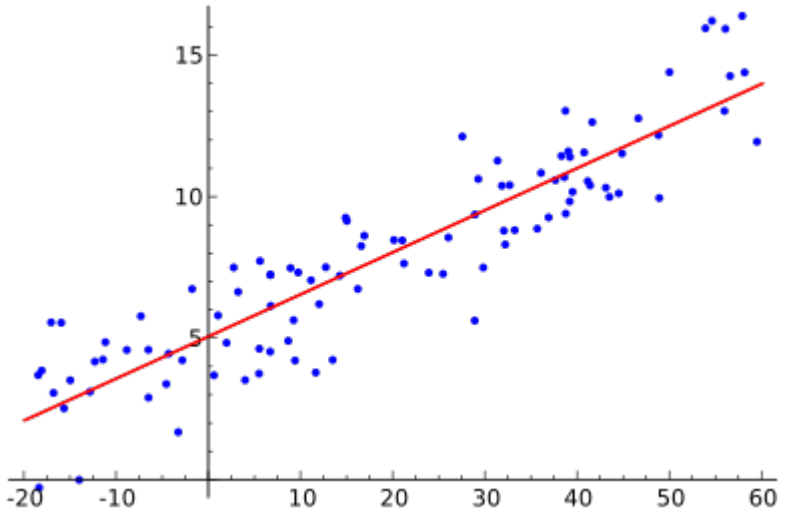


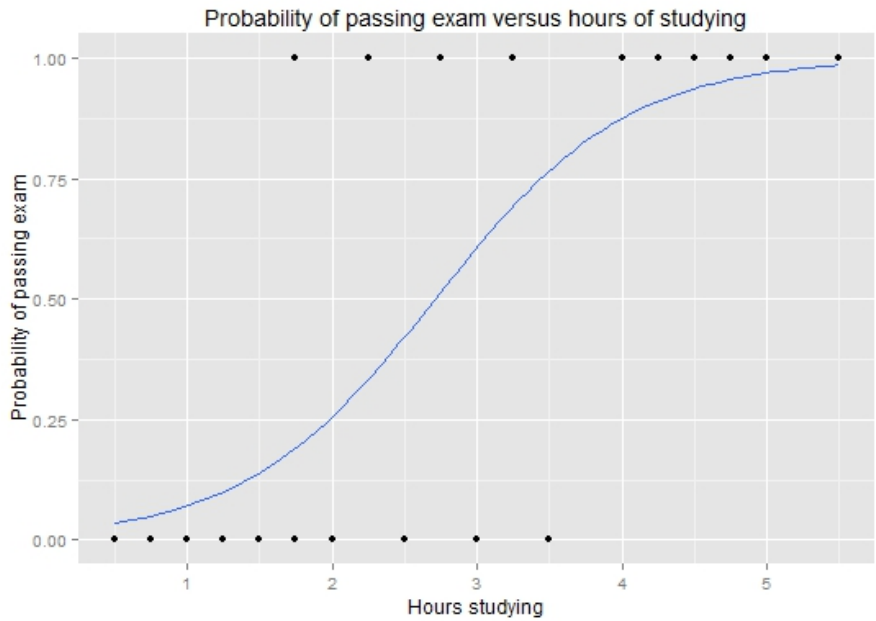
Regression

- Linear (simple or multiple)
- Logistic regression (also called logit)
- Failure time (survival)
- Meta-analysis

- Nonlinear regression

- Feature selection
- Sparse regression





https://en.wikipedia.org/wiki/Regression_analysis

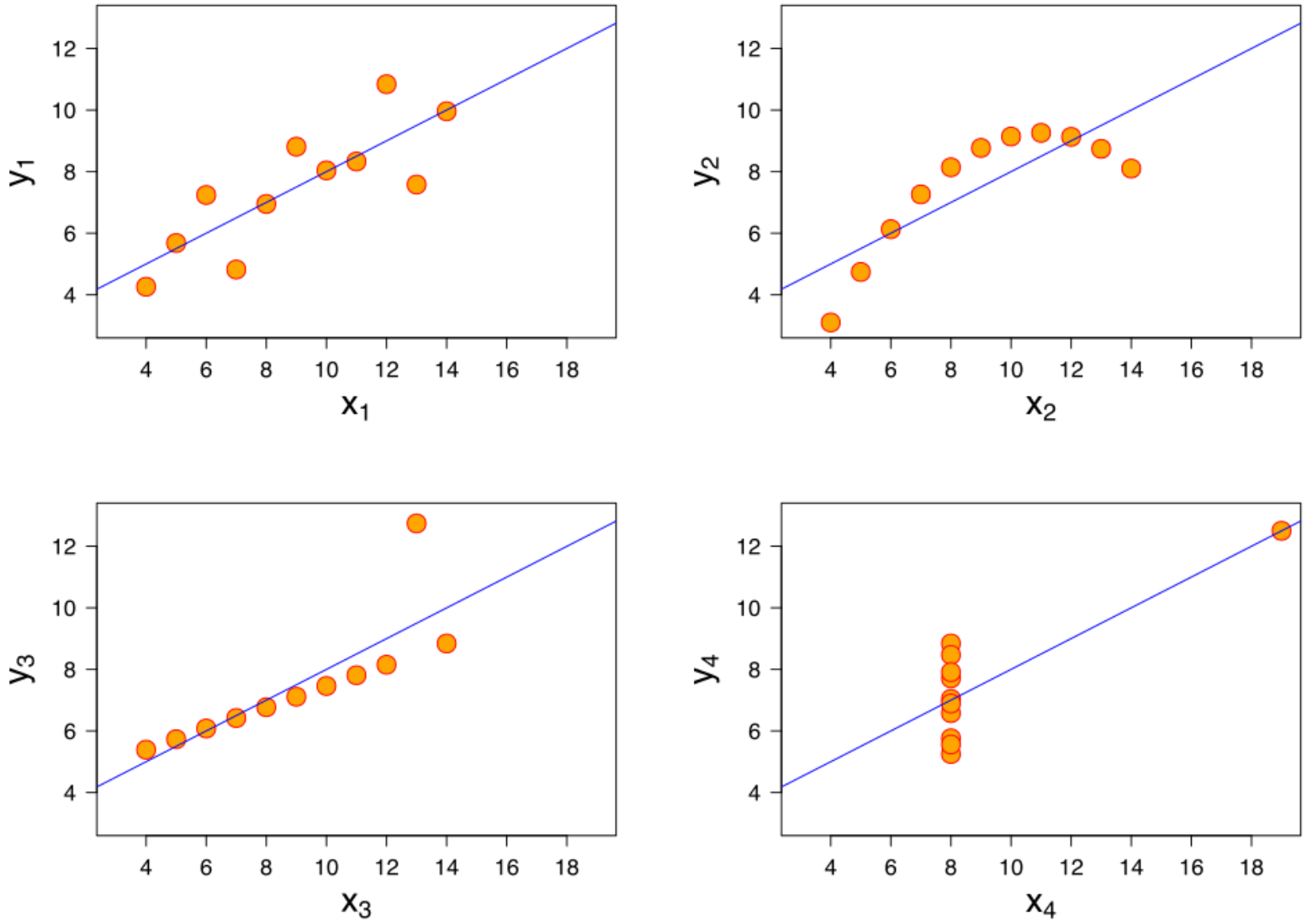
Linear regression

- $Y_i = \beta 0 + \beta_1 x_1 + \beta_2 x_2 + \ldots. + \varepsilon$
- Minimize objective function (sum of squares)
- Linear refers to the coefficients - the regression could be a function of e.g. square of independent variables

Pitfall of linear regression

Same regression line fit to all of these

(but different goodness of fit – plot of residuals)



Anscombe quartet

Linear modeling in R

- lm !

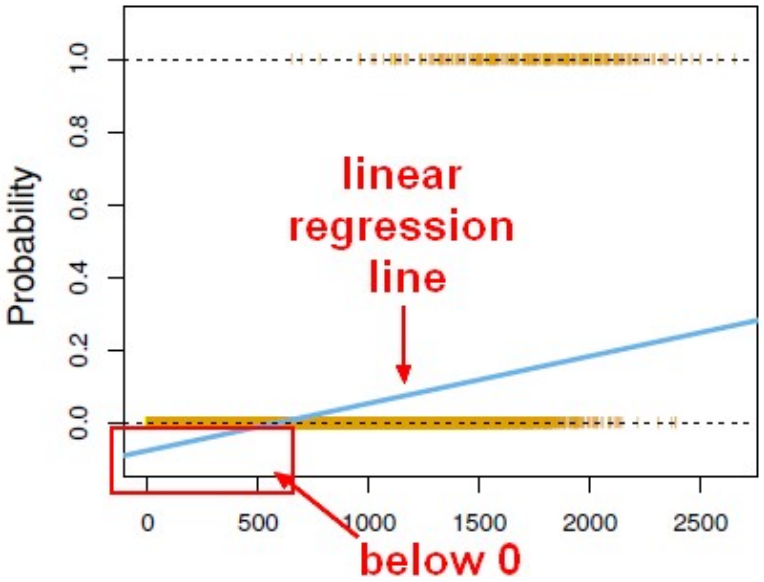
- Two vectors x,y
  - lm(x ~ y)

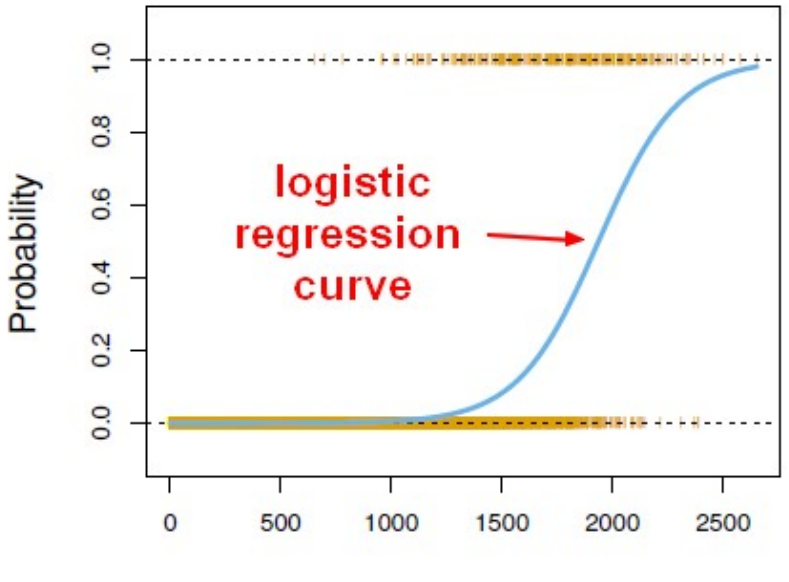
- Two variables in a data frame
  - lm(var1 ~ var2, data=dataframe)

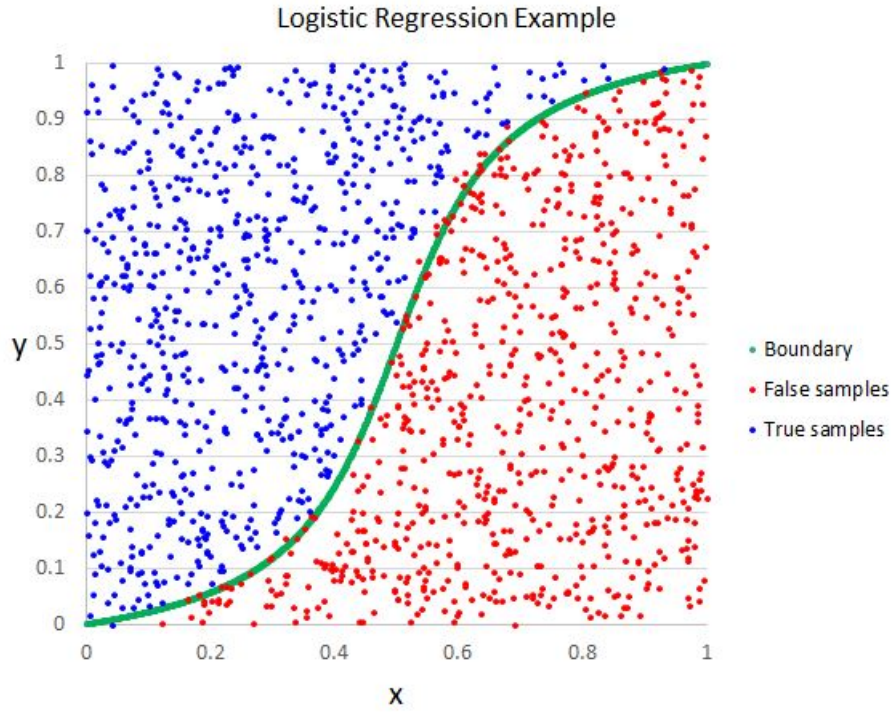
- Residuals = deviation of predicted from observed
- Leverage = ability of an observation to move the regression line

Logistic regression

- Continuous independent variables
- Discrete dependent variable







Survival (Cox) regression

- Time to event (death, relapse, metastasis…)
- Censored – patients drop out of followup

Variable	Patient1	Patient2	Patient3	Patient4	Patient5	Patient6	Patient7	…
Time	12	3	27	8	35	14	22	…
Status	1	1	0	1	0	0	1	…
Gene1	1.45	0.15	1.48	-0.59	-1.88	-0.83	-0.26	…
Gene2	0.94	-0.35	1.23	2.66	-0.23	2.09	-0.13	…
Gene3	0.91	-0.32	0.91	-0.82	-0.35	0.86	0.32	…
…								

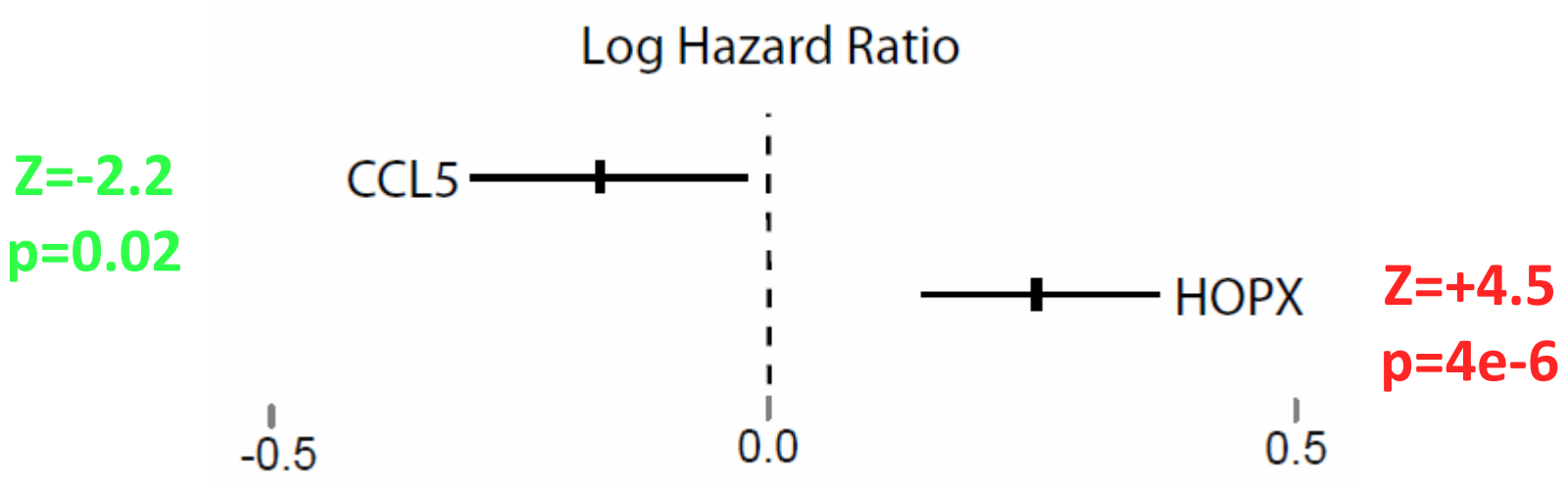
$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

$$h(t, \mathbf{X}) = h_0(t) \times e^{\sum_{i=1}^{p} \beta_i X_i}$$

X_i = expression of gene *i* at sample collection

Expression ~ survival

- Convenient to use the log-hazard

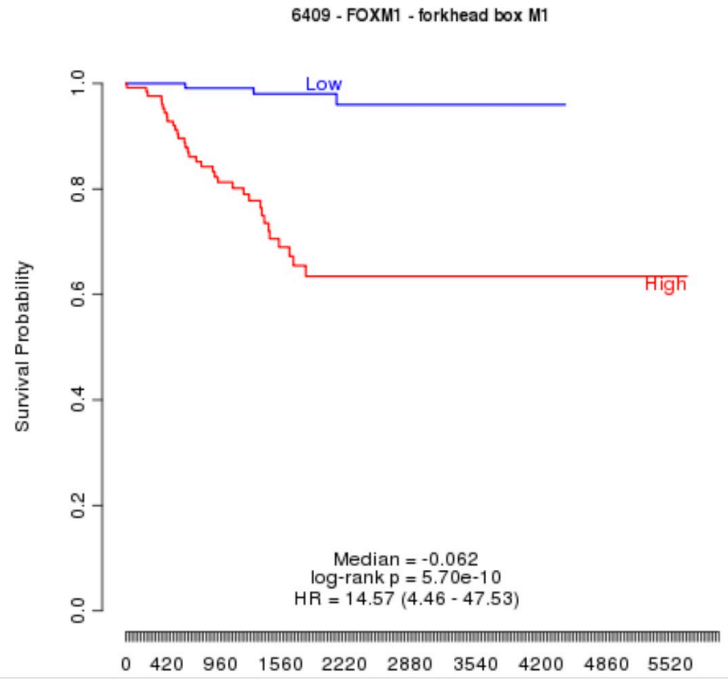


Survival (Cox) regression

- Often evaluated with Kaplan-Meier analysis

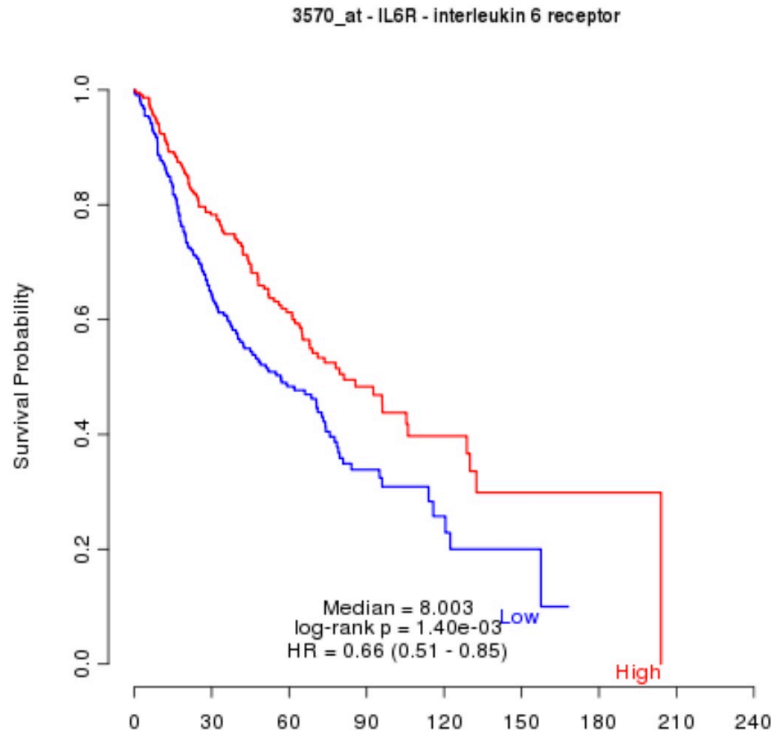
KM Plot(s) for gene **FOXM1** in *Brain cancer Neuroblastoma*

PubMed: Westermann et al. (opens in new browser window/tab)
Accession: E-TABM-38 (opens in new browser window/tab)
No. patients (OS/DSS): 251



KM Plot(s) for gene **IL6R** in *Lung cancer ADENO*

PubMed: Shedden et al. (opens in new browser window/tab)
Accession: ca00182
No. patients (OS/DSS): 255



http://precog.stanford.edu

Feature selection

- Situation:
  - Simple outcome (class, treatment response, survival…0
  - 1000s of potential predictors

  - Do not want all of them in model:
    - Overfitting
    - Uninterpretable

- Sparse regression
  - Find a subset of predictors
  - Lasso & elastic net are common approaches

Lasso/elastic net

- LASSO (least absolute shrinkage and selection operator)
- Idea behind both is that the regression is subject to constraints on coefficients (“budget”)

$$\min_{\beta_0,\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2\right\}\ \text{subject to}\ \sum_{j=1}^{p}|\beta_j|\le t.$$

$$\hat{\beta}=\underset{\beta}{\operatorname{argmin}}(\|y-X\beta\|^2+\lambda_2\|\beta\|^2+\lambda_1\|\beta\|_1)$$

glmnet

- Implements Lasso and elastic net for various models including
  - Continuous outcome
  - Logistic
  - Survival
    - Surv(Time, Status)
- Similar format to samr, lm etc
- Uses nfold-internal cross validation to determine the penalization parameters

Internal cross-validation

- Want to avoid over-fitting
- Split the training data into pieces (e.g. 10)
- Learn model on 90% of data and test it on left-out 10%
- Repeat, leaving out each 10%
- Take parameters which minimize the average error on left-out part
- Run cross-validation multiple times
- Refit model with selected variables

Thursday exercise

- Predict groups (long term vs short term survivors) from cancer data
- Estimate survival function
- Plot survival curves (Kaplan-Meier)

- Please go back and look at some of the genes that came up in Module 1