Tests of homogeneity

Two-tailed test of homogeneity

- Two-tailed test
  - $H_0: m_1 = m_2$
- Principle of the test
  - Estimate the difference between m_1 and m_2
  - Compare this estimation with the theoretical distribution
- Usually, the variance is a priori not know, and has to be estimated
  - Warning: the variance of a difference is the sum of variances
  - The formula for estimating the whether the populations are supposed to have or not similar variances
- The theoretical distribution is thus the *Student (t)*
  - $k = n_1 + n_2 - 2$ degrees of freedom
  - α is shared between the two tails → use the value for $t_{1-\alpha/2}$ in Student's table

$$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\hat{\sigma}_{m_1 - m_2}}$$

Reject H_0 if $t_{obs} \geq t_{1-\alpha/2}$

Homogeneity of a difference

- The test of homogeneity can be thought of as a test of conformity on the difference between two means.
  - H_0: $m_1 = m_2$ $\qquad\qquad$ H_0: $d = |m_2 - m_1| = 0$
- This requires an estimation of the variance of the difference between the two means.
  - The *variance of a difference* between two distributions is the sum of the variances.
  - The *standard error* (i.e. the variance of a the sampling distribution of the mean) is variance of the corresponding population, divided by the sample size.
  - When the variances of the two populations are known *a priori*, the two formulae can be combined to estimate the variance of a difference

Variance of a difference

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Standard deviation of a difference

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Estimating the variance of the difference between sample means

- Generally, the variance of the populations ($\sigma 1$ and $\sigma 2$) are not known a priori.
  - They have thus to be estimated from the two samples.
  - The variance of the sample is a biased estimation of the variance of the population (see chapter on estimation).
  - Each variance estimate needs thus to be corrected by a factor *n/(n-1)*.
- The estimation of the variance will raise an error, which has to be taken into account for the calculation of significance. This will be done differently depending on two considerations
  - Can we assume that the two populations have the same variance ?
  - Do the two sample have the same size ?

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\hat{\sigma}^2_{\bar{X}_1} + \hat{\sigma}^2_{\bar{X}_2}} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

Populations with the same variance

- When one can assume that the two populations have the same variance, the variance of the difference is estimated as follows.

$$\widehat{\sigma}^2 = \widehat{\sigma}_1^2 = \widehat{\sigma}_2^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

$$\widehat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}} = \sqrt{\widehat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

If the two samples have the same size ($n_1 = n_2 = n$), this formula can be simplified.

$$\widehat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{n s_1^2 + n s_2^2}{n + n - 2} \left(\frac{1}{n} + \frac{1}{n} \right)} = \sqrt{\frac{s_1^2 + s_2^2}{n - 1}}$$

Population with different variances

- When one cannot assume that the two populations have the same variance, the variance of the difference is estimated as follows

$$\widehat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{n_1 s_1^2}{n_1(n_1-1)} + \frac{n_2 s_2^2}{n_2(n_2-1)}} = \sqrt{\frac{s_1^2}{n_1-1} + \frac{s_2^2}{n_2-1}}$$

Unequal variances, large samples

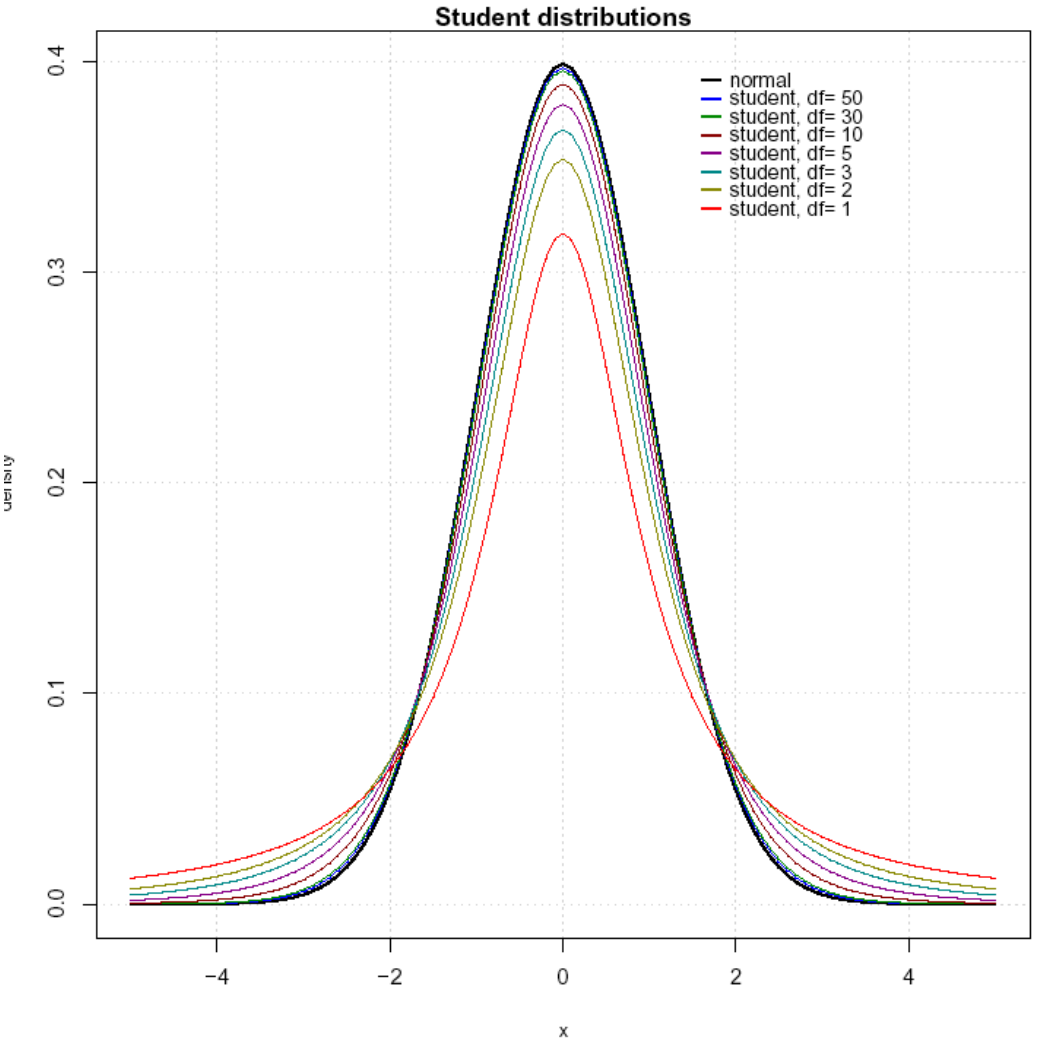
- ***When the two samples are large*** ($n_1 > 30$ and $n_2 > 30$), the Student distribution converges towards a normal distribution.
- The significance of the difference between two means can be assessed with the normal distribution.
- u_{obs}
  - represents the difference between sample means, relative to the estimated standard deviation of this difference.

$$u_{obs} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$$

Equal variances : the Student t-test

- When the two samples not sufficiently large (n1 <30 or n2<30) the same statistics is calculated, but it has to be compared to the Student distribution.
- This test is called ***Student t-test***.
- The shape of the Student distribution depends on one parameter: the degrees of freedom (*k*).
  - Since we assume equal variance, we estimate two parameters for this test: the mean of the difference + the pooled variance.
  - The degrees of freedom are thus
    - $k = n1 + n2 - 2$

$$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\widehat{\sigma}_{\bar{X}_1 - \bar{X}_2}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$$



Unequal variances : the Welch test

- If one cannot assume variance equality, the same statistics (t_{obs}) can be used, but the number of degrees of freedom k is calculated with the formula besides.
  - Note: the formula to compute k in a Welch t-test returns positive Real numbers. The “number” of degrees of freedom does not need to be a Natural number anymore.
- This test is called the ***Welch t-test***.

$$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\widehat{\sigma}_{\bar{X}_1 - \bar{X}_2}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$$

$$k = \frac{\left[\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}\right]^2}{\frac{1}{n_1 - 1}\left[\sqrt{\frac{s_1^2}{n_1 - 1}}\right]^2 + \frac{1}{n_2 - 1}\left[\sqrt{\frac{s_2^2}{n_2 - 1}}\right]^2}$$

Student or Welch ?

- When searching for differentially expressed genes, should we apply Student or Welch test ?
- In transcriptome analysis, we generally assume that the variance will be somewhat proportional to the expression level.
- We can thus not assume equality of variance between low- and high-expressing conditions, respectively.
- The Welch test is thus a priori more appropriate to detect differentially expressed genes in transcriptome array profiles.

Selection of differentially expressed genes

- The test of homogeneity can be applied to select genes differentially expressed between two experimental conditions, cell types, …
- Example: Golub data
  - Oligonucleotide arrays were used to measure the level of expression of > 7000 genes in
    - 27 patients suffering from acute lymphoblastic leukemia (ALL)
    - 11 patients suffering from acute myeloblastic leukemia (AML)
  - An *ad hoc* preliminary filtering was done by the authors, leaving 3051 genes, considered as reliable measurements.
  - **Question:** which genes show a significantly different level of expression between the two patient types (ALL and AML, respectively) ?
  - **Approach:** apply the test of homogeneity to each gene g separately, to test the null hypothesis H_0: $m_{g,ALL}=m_{g,AML}$

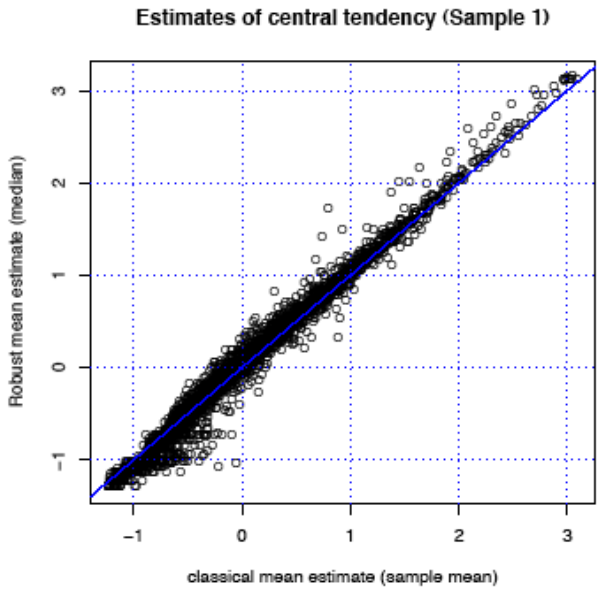
Multiple student test

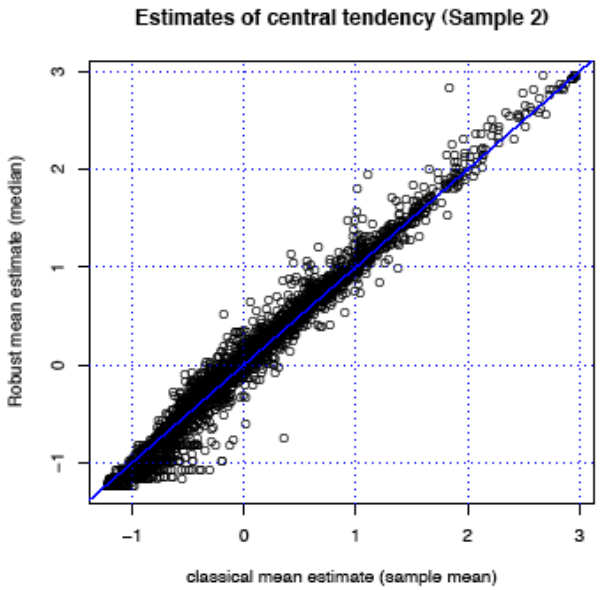
- Goal : select genes differentially expressed between distinct patient types
- Method:
  - 2 patient types : T-test
    - Assumption of equal variance is generally not valid -> use Welch test instead of Student test.
  - >2 patient types: ANOVA (not shown here)
- **Attention : problem of multi-testing**
  - We are testing several thousands of probes in parallel.
    - 12,578 for response to X (data from Thierry Lequerre)
    - 3050 for the cancer type (Golub 1999)
  - If we accept the “conventional” alpha risk of 0.01, we expect
    - 126 false positives for the response to X
    - 30 false positives for the cancer type

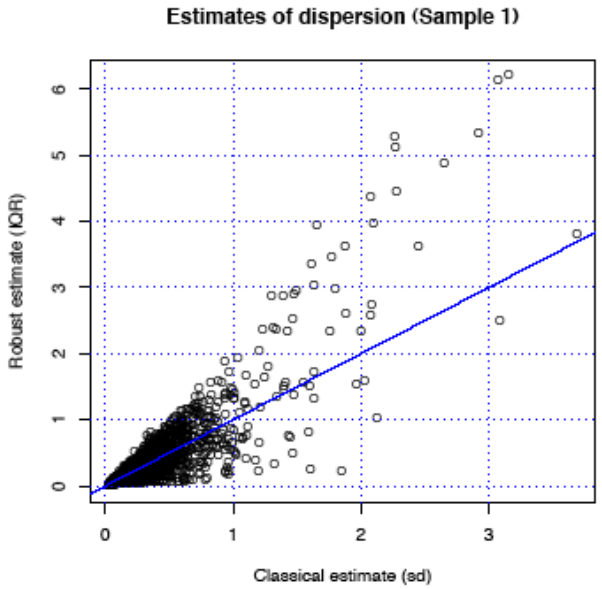
Multi-testing corrections

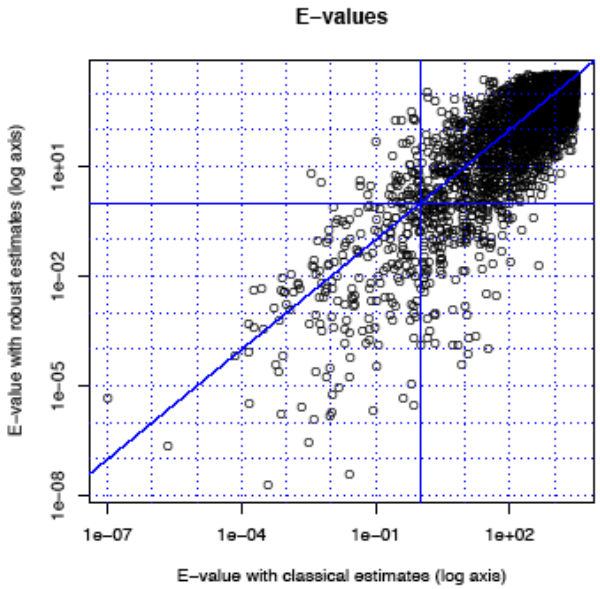
- Bonferoni rule
  - Reduce alpha risk according to the number of tests (*T*)
  - $P\text{-value} < 1/T$
- E-value (equivalent to Bonferoni rule)
  - Estimate the expected number of false positives
  - $E\text{-value} = T * P\text{-value}$
- Family-Wise Error Rate (FWER)
  - Probability to observe at least one false positive in T tests.
  - $FWER = 1 - (1 - P\text{-value})^T$

Choice of the estimators for the t-test



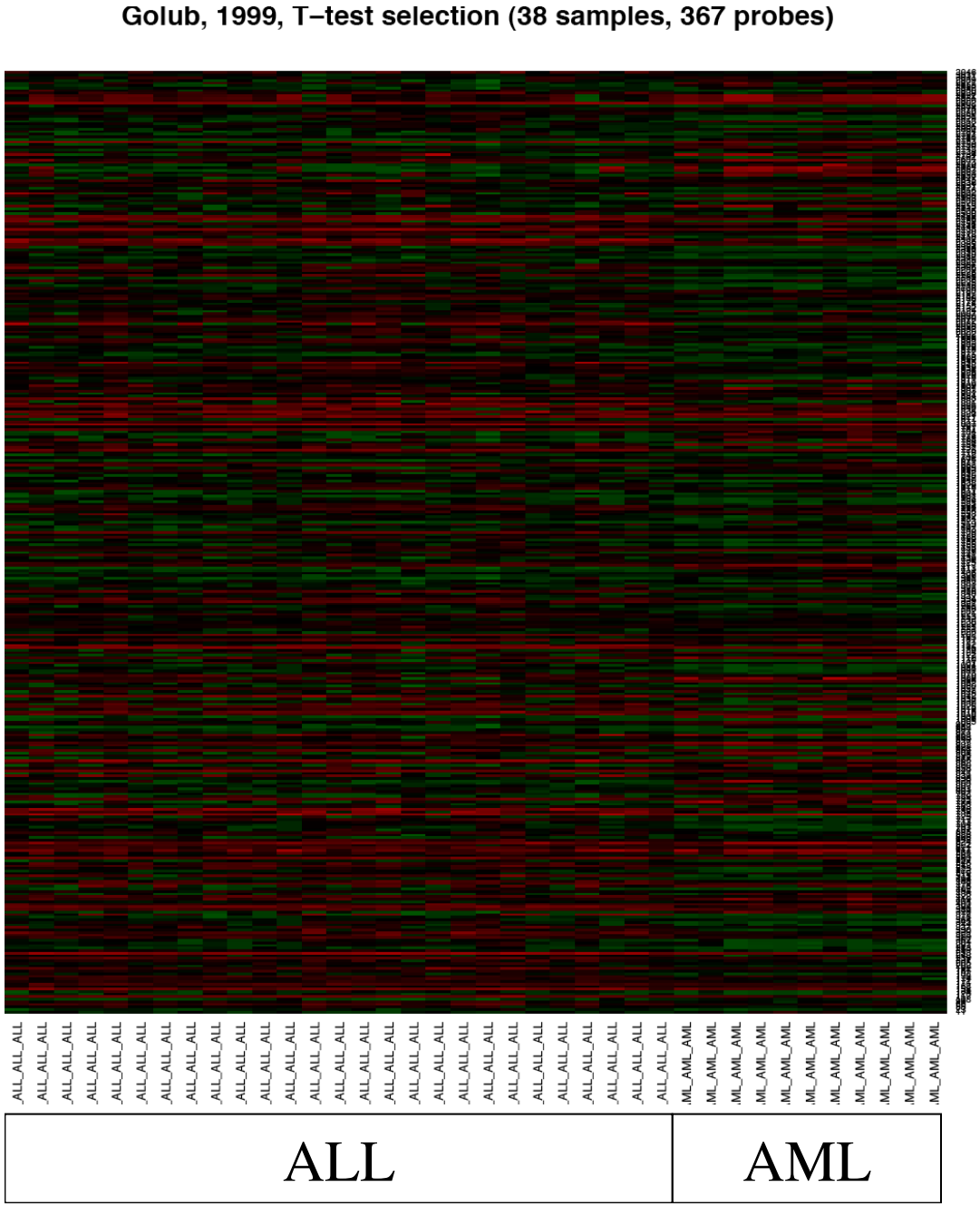






- A priori, I would recommend use robust estimators not only for standardization, but also for the t-test.
  - central tendency (median instead of mean).
  - Dispersion: IQR instead of standard deviation.
- The most significant genes are the same irrespective of this choice, but for some other genes it makes a change.
  - Classical estimators: 243 significant genes
  - Robust estimators: 367 significant genes
  - Genes selected in both cases: 197.

Golub 1999 - Profiles of selected genes



- The 367 gene selected by the T-test have apparently different profiles.
  - Some genes seem greener for the ALL patients (27 leftmost samples)
  - Some genes seem greener for the AML patients (11 rightmost samples)
- However, the relationships between the different genes is not visible on this map.
  - In the next courses, we will use clustering methods in order to regroup genes with similar profiles, among those selected as differentially expressed.