

Lecture 6: Two-group comparative experiments

Objective: compare the expression levels of a set of genes across two or more conditions

Find genes that are significantly differentially expressed across conditions.

Example: Compare GEL in cancerous liver cells versus normal liver cells.

Example: compare GEL in cancerous liver cells treated with a drug Vs control

Simplest analysis: Each gene in isolation - compare GEL's across groups

Complex analysis: Genes are analyzed in combination,
 comparing the expression levels of clusters of genes across the groups.
 collect genes differentially expressed across the groups
 deduce regulatory pathways

Example: Comparing gene expression profiles (GEP) of two groups of four mice:(4 control, 4 treated with a test compound). Several hours post-treatment, an mRNA sample was extracted from the liver of each animal and placed on a microarray containing 4077 genes and expression levels were obtained. Data is log transformed and normalized. See scatterplot matrix.

```
X = data.compound
min(X)
pre.summary(X)
## The choice is k = 0
Y = f.qn(log(X))
f.pairs(Y)
f.concor(Y)
f.concor.map(Y)
```

Notation: n_1 microarrays in Group 1 and n_2 microarrays in Group 2; the total sample size is $N=n_1+n_2$. x_{gij} : intensity measurement for the g th gene in the i th microarray in the j th group, where $i=1,\dots,n_j$, $j=1,2$; and $g=1,\dots,G$.

$\bar{x}_j, \tilde{x}_j, s_j, \tilde{s}_j$ denote, respectively, the mean, median, standard deviation and median absolute deviation from the median (mad) of the j th group.

Basics of statistical hypothesis testing

null hypothesis there is no difference between groups

G null hypotheses being tested, one per gene.

Decision: positive finding: reject the null hypothesis and claim that there is a difference between the groups

negative finding: do not reject the null hypothesis and declare there is insufficient evidence to detect a difference between the groups

true positive

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Decision
Do not Reject H_0	Correct Decision	Type II Error

$P(\text{Type I Error}) = \alpha$
 $P(\text{Type II Error}) = 1 - \beta$
 $\beta = \text{Power}$
 $P(\text{Reject } H \text{ when } H \text{ is False}) = \beta$

Reject: true positive, false positive or a Type I error

Not reject: true negative, false negative or a Type II error

test statistic. $T = \frac{r}{s}$. r estimate of the size of the biological effect being

tested; the further the data are from the null the larger the value of r . s is a standard error that measures the variability of r ("signal" vs "noise")

null distribution: Prob distrib of the test statistic under the null hypothesis

p-value: Prob. of observing a value as extreme as that observed if the null hypothesis was true. small *p-value*, strong evidence against the null hypothesis.

significance level : α : (typically 5%) *p-value* < significance level : reject null

The probability of the test reaching a false positive decision is called the

$P(\text{false positive rate}) = P(\text{Type I error})$

$P(\text{false negative rate})$ (or the *Type II error probability*). Also, the

specificity = $P(\text{"true positive"})$ = true positive rate

sensitivity = $P(\text{"true negative"})$ = true negative rate

Neyman-Pearson approach to statistical hypothesis testing. The false positive rate is controlled at a specified small value, called the *size* of the test, and then the test is set up to have as small a false negative rate (or equivalently as high a true negative rate, which is called the *power* of the test) as possible – in other words, “fix the size, maximize the power”.

Fold changes:

Arabidopsis thaliana, Schena *et al* (1995) five-fold difference

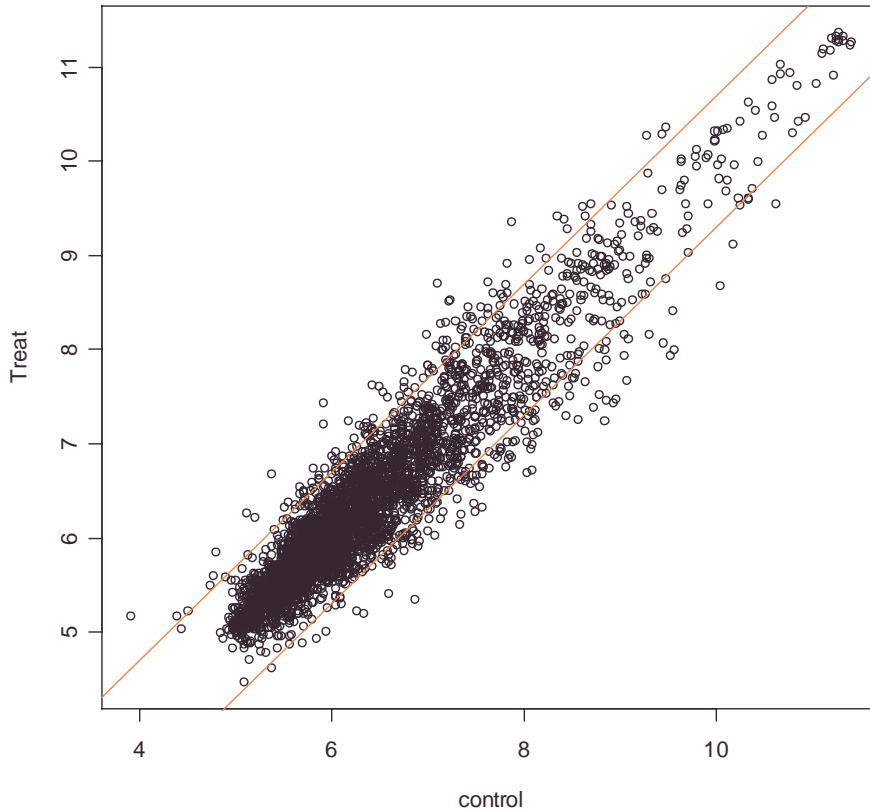
h-fold difference if $|\bar{x}_2 - \bar{x}_1| > \log(h)$

Example: For the example dataset, Figure 7.2 shows a histogram of the $D = \bar{x}_2 - \bar{x}_1$ values, which range from -1.66 (5.26-fold) to 1.72 (5.58-fold)

Near zero median of 0.01 with

119 two-fold or greater upregulation

144 genes showing a two-fold or greater downregulation



The two-sample t test

two-sample t test statistic $T_e = \frac{|\bar{x}_1 - \bar{x}_2|}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$.

If the data are drawn from a Gaussian distribution (the Gaussian distribution is sometimes called the normal distribution) and homoscedastic (*i.e.*, have equal variances): $x_{ij} \sim N(\mu_j, \sigma^2)$, The null distribution of T_e is a t -distribution with degrees of freedom $\nu = n_1 + n_2 - 2$. If the observed value of T_e is $T_{e;obs}$ then the p -value is given by the probability $p_e = \text{Prob}(|T_e| > T_{e;obs})$. A gene is declared significantly differentially expressed at level of significance α if $p_e < \alpha$.

Example:

Observe that the t statistic has the form of a signal-to-noise ratio as mentioned in Section 7.1. The "signal" is the numerator that reflects the

difference we are trying to find; the “noise” is the denominator that reflects the variability of the system.

Non homoscedastic errors: Welch’s test statistic

$$T_u = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The null distribution of T_u is, approximately, a t -distribution with degrees of freedom:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{(n_1 - 1)}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{(n_2 - 1)}\left(\frac{s_2^2}{n_2}\right)^2}$$

If the observed value of T_u is $T_{u,obs}$ then the p -value is given by the probability $p_u = \text{Prob}(|T_u| > T_{u,obs})$. A gene is declared significantly differentially expressed at level of significance α if $p_u < \alpha$.

Diagnostic checks

The *residuals*, $r_{ij} = x_{ij} - \tilde{x}_j$, data follow a Gaussian distribution.

Gaussian probability plot,

the *standardized residuals*, $r_{ij}^* = \frac{x_{ij} - \tilde{x}_j}{\tilde{s}_j}$, may be used instead.

Again a resistant measure of scale, \tilde{s}_j , is used instead of the traditional measure of scale, s_j . Large absolute values of r_{ij}^* indicate outliers.

With microarray data, it is in fact useful to gather all the residuals across all the genes $\{r_{gij}^*\}$ for making the Gaussian probability plot.

Example:

This graphical check is often enough, but there are several formal statistical tests for assessing Gaussianity as well. One of the most effective is the Shapiro-Wilk test. Other tests include the Kolmogorov-Smirnov test and its modifications, such as the Anderson Darling test. With a very large number of observations, however, these tests will indicate nonGaussianity even with

trivial departures from perfect Gaussianity. Therefore we shall not use them here.

Formal tests for unequal variances across groups, such as Bartlett's test and Levene's test, require larger sample sizes

Plot $\{s_{gi1}^{2/3}\}$ versus $\{s_{gi2}^{2/3}\}$. (Wilson and Hilferty (1931))

Example:

Robust t -tests: The t -test can be rendered *robust* by replacing the means and variances in the test statistic with robust versions of these sample statistics.

Randomization tests:

Use permutations of the columns of the GEM

Example: In the example, there are two groups of four, making for 35 possible permutations.

We use the difference in means, $T_d = \bar{x}_2 - \bar{x}_1$, as test statistic and will regard T_d as significant if the observed value $|T_{d;obs}|$ of $|T_d|$ exceeds $|T|$ in at most one permutation, which constitutes a two-sided test of level $2/35=5.7\%$, which we will call 5% without quibbling over the extra 0.7%.

Note that a randomization test is robust to outliers only if the test statistic itself is resistant.

The Mann-Whitney-Wilcoxon rank sum test

When it is clear that the underlying distribution is far from Gaussian, it may still be reasonable to assume that the distributions for the two groups are identical except for a location effect, so that $X_{i1} \sim F(\mu)$, $X_{i2} \sim F(\mu + \theta)$, where $F(\mu)$ denotes a distribution centered at μ . The Mann-Whitney-Wilcoxon test can be used to test the hypothesis that location parameter $\theta=0$.

Once the observations have been ranked from 1 to N in increasing order, the test statistic for the Mann-Whitney-Wilcoxon test is the *rank sum statistic*, R , the sum of the ranks corresponding to the observations in Group 1. This

statistic measures the degree of overlap between the two groups, the smaller the overlap, the further the value of R is from its null value of $n_1(N+1)/2$, indicating a group difference.

The null distribution of R has been tabulated (see, *e.g.*, Hollander and Wolfe (1999)) for small values of n_1 and n_2 using an argument similar to that of permutation tests. For larger values of n_1 and n_2 , the fact that

$$\frac{|R - \frac{n_1(N+1)}{2}|}{\sqrt{\frac{n_1 n_2 (N+1)}{12}}}$$

has, approximately, a standard Gaussian distribution under the null hypothesis can be used to obtain p -values. If the observed value of R is R_{obs} then the p -value is given by is the probability $p_R = \text{Prob}(|R| > R_{obs})$. A gene is declared significantly differentially expressed at level of significance α if $p_R < \alpha$.