Microarray Workshop

t-test and other related tests

München, Februar 2004

Shaun Seaman

MPI für Psychiatrie

Overview

- Example dataset

- Score test

- Two-sample t-test

- ANOVA

- Mann-Whitney two-sample / Wilcoxon Rank Sum test

- Kruskal-Wallis test

Example Dataset

Twenty mice, 10 from strain 1, 10 from strain 2.
Five mice from each strain receive treatment 1.
Five mice from each strain receive treatment 2.

Twenty identical microarrays.
Simple indirect comparison design.
Label treatment RNA with green dye.
Label common reference with red dye.
Co-hybridise treatment RNAs and common reference to 20 arrays.

Consider just one probe (there are thousands on the whole array).
The 20 $\log_2(G/R)$ measurements are:

Treatment 1		Treatment 2	
Strain 1	Strain 2	Strain 1	Strain 2
-1.077	-3.023	-3.047	-0.463
0.805	0.221	-4.639	-1.039
0.457	-1.522	-2.336	-2.050
4.047	-1.340	-1.119	-3.819
5.149	0.549	0.760	-4.668

Questions

1. Is expression different in two treatment groups?

2. Is expression different in two strain groups?

3. Is there a treatment-strain interaction?

4. Pretend four treatment-strain groups are actually four treatment groups.
   Is expression different in these four treatment groups?

Score test

(Comparing mean in two groups)

Let X_{ij} $(j = 1, \ldots, n_i)$ be expression in mouse j of group i $(i = 1, 2)$.

Assume

$$X_{ij} = \mu_i + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$.

Null hypothesis is $H_0 : \mu_1 = \mu_2$.

Let $\bar{X}_i$ be mean expression in group i.

Let $\bar{X}$ be mean expression overall (in groups 1 and 2).

$$
\begin{aligned}
\bar{X}_1 &= \frac{\left(\begin{array}{c} -1.077 + 0.805 + 0.457 + 4.047 + 5.149 \\ -3.023 + 0.221 - 1.522 - 1.340 + 0.549 \end{array} \right)}{10} \\
&= 0.427 \\
\bar{X}_2 &= -2.242 \\
\bar{X} &= -0.908
\end{aligned}
$$

So, $\bar{X}_1 - \bar{X}_2 = 2.669$.

If H_0 is true,

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0, 1)$$

But we do not know σ^2.

Estimate σ^2 by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

Then

$$\frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{1/n_1 + 1/n_2}} \sim N(0, 1)$$

approximately.

With example data,

$$
\begin{aligned}
s^2 &= \frac{1}{20 - 1} \left[\begin{array}{l} (-1.077 - -0.908)^2 + \cdots \\ +(-4.668 - 0.908)^2 \end{array} \right] \\
&= 6.421
\end{aligned}
$$

So,

$$\frac{2.669}{\sqrt{6.421 \left(\frac{1}{10} + \frac{1}{10} \right)}} = 2.355$$

Now, if $Z \sim N(0, 1)$, then $P[|Z| > 2.355] = 0.0185$.

So, approximate p-value is 0.019.
We reject H_0 at 5% level.

Two-sample t-test

Score test sometimes used, but t-test better. More powerful and gives exact p-value.

Instead of using

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

Use 'pooled' estimate

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

where s_i^2 is estimate of σ^2 based on data from group i:

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

If H_0 true,

$$\frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1 + n_2 - 2}$$

exactly.

For example data, $s_p^2 = 4.799$

$$\frac{2.669}{\sqrt{4.779 \left(\frac{1}{10} + \frac{1}{10}\right)}} = 2.702$$

If $T \sim t_{18}$, then $P[|T| > 2.702] = 0.0146$.

So, exact p-value is 0.0146.

ANOVA (Analysis of Variance)

Two-sample t-test is for two groups.
ANOVA is generalisation to more than two groups.

One-way ANOVA

Assume

$$X_{ij} = \mu_i + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ $(i = 1, \ldots, I;\ j = 1, \ldots, n_i)$.

Null hypothesis is $H_0 : \mu_1 = \ldots = \mu_I$.

Look at ratio of variation between groups to variation within groups. If ratio is large, suggests H_0 is false.

Formally,

$$B = \frac{\sum_{i=1}^{I} n_i (\bar{X}_i - \bar{X})^2}{I - 1}$$

is between-group variation. Within-group variation is

$$W = \sum_{i=1}^{I} \frac{(n_i - 1)s_i^2}{n - I}$$

If H_0 true, B/W has F-distribution with $(I - 1)$ and $(n - I)$ degrees of freedom.

For example dataset, $I = 4$ and

$$
\begin{aligned}
B &= \frac{5}{4-1}[(1.876 - -0.908)^2 + (-1.023 - -0.908)^2 \\
&\quad + (-2.076 - -0.908)^2 + (-2.408 - -0.908)^2] \\
&= 18.96
\end{aligned}
$$

$$
W = \frac{5-1}{20-4}(6.827 + 2.092 + 4.135 + 3.221) = 4.07
$$

So, $B/W = 4.66$.

If $F \sim F_{3,16}$ then $P[F > 4.66] = 0.0159$.
So, p-value is 0.016.

One-way ANOVA with two groups gives identical p-values to two-sample t-test

Two-way ANOVA

Suppose there are two factors, rather than just one, e.g. treatment and strain. Assume

$$X_{ijk} = \alpha_i + \beta_j + \epsilon_{ijk} \tag{0.1}$$

where $\epsilon_{ijk} \sim N(0, \sigma^2)$
$(i = 1, \ldots, I, \; j = 1; \ldots, J; \; k = 1, \ldots, n_{ij})$.

E.g. $\alpha_1, \ldots, \alpha_I$ treatment effects and $\beta_1, \ldots, \beta_I$ strain effects.

Can test null hypothesis of no treatment effect: $H_0 : \alpha_i = \alpha$ for all i; or of no strain effect: $H_0 : \beta_j = \beta$ for all j.

Interaction

Equation (0.1) is for model without interaction, i.e. effect of treatment assumed same for all strains. Interaction means effect of treatment differs according to strain. Adding interaction terms gives

$$X_{ijk} = \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

Test of no interaction is test of null hypothesis that all γ_{ij}'s equal zero.

ANOVA extends to more than two factors.

Mann–Whitney two-sample / Wilcoxon Rank Sum test

T-test assumes normally distributed errors and equal error variances in two groups.
A variant exists (Welch test) that allows different error variances, but it still assumes normality.

When data not normally distributed, type-I error rate higher than nominal level. Too many false positives.

Mann–Whitney test (a.k.a. Wilcoxon Rank Sum test) is 'non-parametric', i.e. does not make normality assumption. It maintains correct type-I error rate.

Test works with ranks of expression data, rather than actual values. Ranks of example data are

Treatment 1		Treatment 2	
Strain 1	Strain 2	Strain 1	Strain 2
11	5	4	13
18	14	2	12
15	8	6	7
19	9	10	3
20	16	17	1

Ranks for treatment 1 tend to higher than those for treatment 2.
Sum of ranks in group 1 is $R = 11+18+\ldots+16 = 135$.
Total sum of ranks is $1 + 2 + \ldots + 20 = 210$.
If H_0 is true, would expect R to be about $210/2 = 105$.
Is $R = 135$ unlikely under H_0?

Under H_0, all permutations of ranks equally probable.

16	14	9	7
11	4	13	18
12	2	17	20
5	19	15	10
8	1	6	3

$R^* = 92$

6	15	3	20
9	12	8	18
16	19	10	14
4	17	13	2
1	11	5	7

$R^* = 110$

4	1	5	3
14	13	16	10
8	18	17	2
12	20	7	19
9	11	6	15

$R^* = 110$

Exact p-value is obtained by looking at all $\binom{20}{10} =$ 6.7×10^9 possible permutations.
Calculate R^* for each permutation.

P-value is the proportion in which $R^* > 135$ multiplied by two.

6.7×10^9 is very large number to evaluate, but can randomly generate, say, 10,000 permutations and estimate p-value.

Also, for large samples, distribution of R under H_0 can be approximated and an approximate p-value calculated.

For example dataset, $p = 0.023$.

Wilcoxon test is *robust*, but less powerful when normal assumption is true.
(In this example, $p = 0.023$ compared to 0.0146 for t-test.)
Loss of power quite small for large samples. Can be large for small samples.

Note: Wilcoxon test is example of a permutation test. These are common in statistics. All use principle that under H_0 all permutations of the data are equally probable. Calculate some statistic (e.g. R) and see in what proportion of permutations the statistic is as extreme as value actually observed.

Kruskal-Wallis test

Kruskal-Wallis test is generalisation of Wilcoxon test to more than two groups.

Non-parametric equivalent of one-way ANOVA.

Does not assume normally distributed data, and so is robust.

However, if data are normally distributed, ANOVA is more powerful.

For the example dataset, the Kruskal-Wallis test for equality of mean expression in the four groups gives $p = 0.048$ (compared to 0.016 for ANOVA).