Lecture 4: Hypothesis testing for continuous variables

BTBI30081

統計應用方法Applied Methods in Statistics

2025/3/12

Example: Gene expression microarray data

- Data from a study using gene expression profiling to predict breast cancer outcomes (http://www.nature.com/nature/journal/v415/n6871/full/415530a.html)
- 78 breast cancer: 44 remained disease-free for an interval of at least five years after their initial diagnosis (good prognosis group), while 34 patients had developed distant metastases within five years (poor prognosis group)

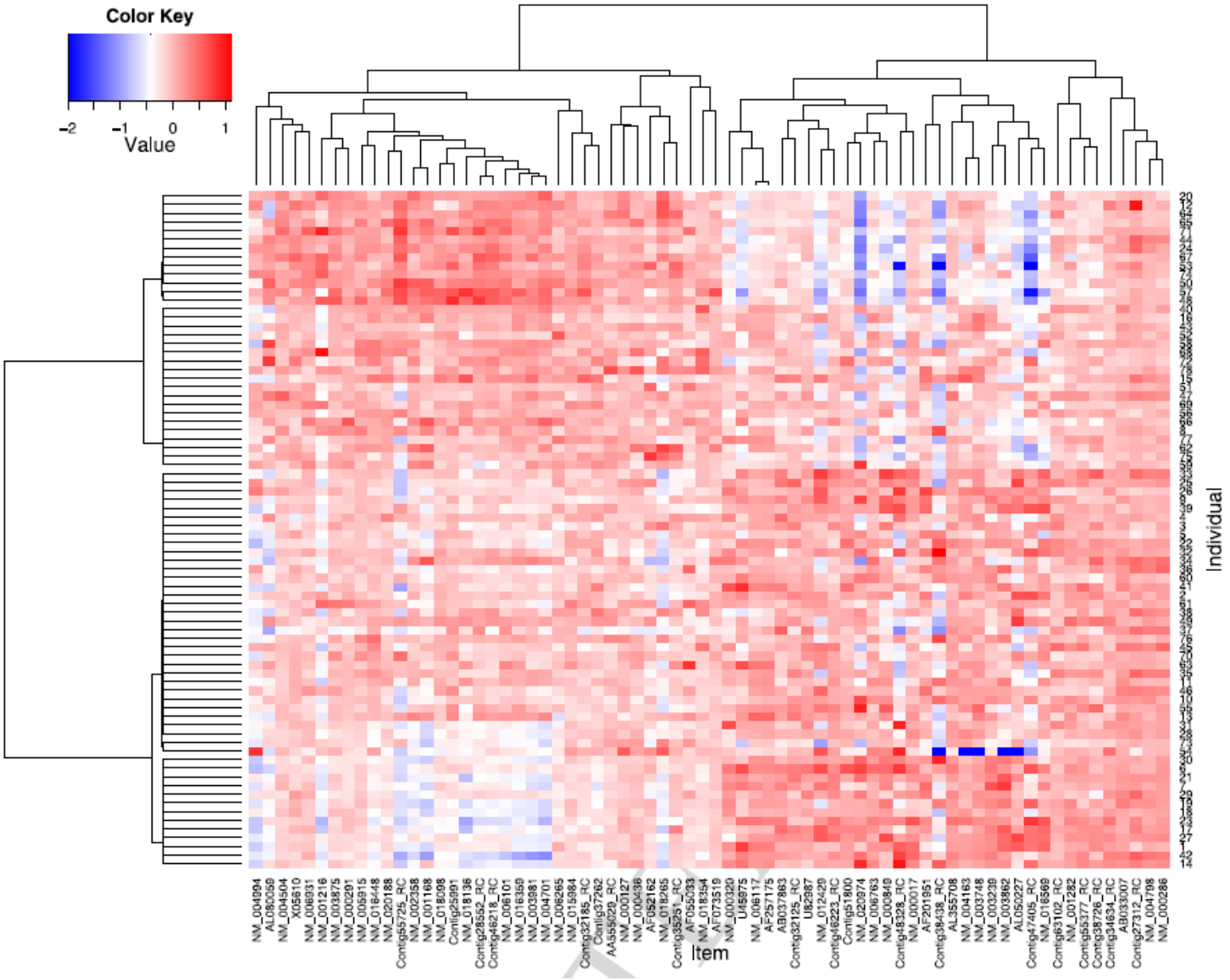
samplexprs.csv

Variable	Description
id	An unique identification number
age	Age at diagnosis of breast cancer (year)
metastases	Developing distant metastases: 0=no (good prognosis group), 1=yes (poor prognosis group)
followup	Follow-up time (year)
ERp	ER-α expression level
J00129	$\log_{10}$ gene expression intensity ratios
Contig29982_RC	$\log_{10}$ gene expression intensity ratios

- RMD_example 4.1

id	age	metastases	followup	ERp	J00129	Contig29982_RC	Contig42854	Contig42014_RC
FG80	52	0	7.35	100	-0.795	-0.387	0.199	-0.247
SF58	50	1	1.15	0	-0.509	0.459	-0.257	-0.065
DE72	54	0	12.12	100	-0.961	-0.631	0.037	-0.153
DE65	40	0	6.25	0	-0.749	0.699	-0.346	0.032
HG87	53	0	5.18	0	-0.426	-0.406	-0.355	0.429
HG88	37	1	1.09	100	-0.566	-0.596	-0.352	-0.336
AB22	37	0	5.8	90	-0.42	-0.286	-0.09	-0.048
HG91	30	1	1.03	0	-0.499	-0.402	0.181	0.143
HG92	39	1	3.36	80	-0.465	-0.533	-0.019	0.019
KH11	45	1	1.62	50	-0.189	-0.309	-0.152	0.918
KH20	30	1	4.7	70	-0.739	0.093	-0.214	-0.025
SF67	48	1	1.98	0	-0.601	-0.177	-0.2	0.108
LD44	33	1	1.4	0	0.786	-0.164	-0.144	0.027
AA04	41	0	13	50	-0.819	-0.267	0.023	-0.23
AA01	43	0	12.53	80	-0.448	-0.296	-0.1	-0.177
GL73	52	1	2.13	0	1.206	-0.353	-0.039	-0.006
AA10	49	0	11.16	80	-0.391	-0.31	-0.06	-0.164
HG86	54	0	5.89	50	-0.234	-0.404	-0.214	0.421
DE62	40	0	6.97	50	-0.75	-0.316	-0.021	-0.041
AB26	41	0	8.17	10	-0.299	-0.137	-0.214	0.031
SF57	41	1	2	0	-0.455	-0.288	-0.241	-0.032
DE61	45	0	13.42	100	-1.173	-0.887	-0.058	0.021

Example: Gene expression microarray data (samplexprs.csv)



Heatmap for gene expression microarray data (samplexprs.csv)

One-sample t-test

- Compare the mean of the sample to a given number
- Perform t-test for testing whether or not the population mean of $\log_{10}$ gene expression intensity ratios on gene J00129 (μ_{J00129}) is equal to -0.5
- *Ho* : $\mu_{J00129} = -0.5$

  Ha : $\mu_{J00129} \neq -0.5$

  (RMD_example 4.2)

Two-sample t-test

- Compare the mean of the first sample minus the mean of the second sample to a given number, where two samples are independent
- Perform t-test for testing whether or not the difference of population mean $\log_{10}$ expression intensity ratios on gene J00129 between good (μ_G) and poor (μ_P) prognosis groups is equal to 0.
- *Ho* : $\mu_G = \mu_P$ *Ha* : $\mu_G \neq \mu_P$ (RMD_example 4.3)

Notes on two-sample t-test

- Only be used if the variances of the two samples are assumed to be equal
- When this assumption is not true, the test used is called Welch's t-test.
- These tests are applied when the statistical units underlying the two samples being compared are non-overlapping (i.e., independent).

F-test for equal variance

- Test whether the variances of the two samples are equal
- Perform F-test for testing whether or not the variance of $\log_{10}$ expression intensity ratios on gene J00129 for the good prognosis group (σ_G) is equal to the variance for the poor prognosis group (σ_P)
- *Ho* : $\sigma_G = \sigma_P$ *Ha* : $\sigma_G \neq \sigma_P$ (RMD_example 4.3)

ANOVA

- Compare the means among more than two samples, where all samples are independent and the variances of these samples are equal
- ANOVA for testing the equality of population mean $\log_{10}$ expression intensity ratios on gene J00129 among 11 ERp groups (0, 5, 10, 30, 40, 50, 60, 70, 80, 90, 100)
- *Ho* : $\mu_0 = \mu_5 = \mu_{10} = \mu_{30} = \mu_{40} = \mu_{50} = \mu_{60} = \mu_{70} = \mu_{80} = \mu_{90} = \mu_{100}$

  Ha : not *Ho*

  (RMD_example 4.4)

Paired t-test

- Compare the difference between two responses measured on the same statistical unit
- Perform paired t-test for testing whether or not the difference of the $\log_{10}$ expression intensity ratio on gene J00129 and the one on Contig29982_RC *from the same individual* is equal to 0
- *Ho* : $\mu_{J00129} = \mu_{Contig29982\_RC}$

  Ha : $\mu_{J00129} \neq \mu_{Contig29982\_RC}$

  (RMD_example 4.5)

Notes on paired t-test

- Here, the two samples under comparison are not independent. They are from the same unit, and are correlated.

Notes on above t-tests

- Data are assumed to be normally distributed.
- Or, the sample size needs to be large enough.

Permutation test

- Permutation method can be an extension from any powerful testing methods.
- The null distribution is calculated by randomly permuting (shuffling) the class labels.
- We can the estimate the p-value of a test.
  - How many of the null means are bigger than the observed value? That proportion would be the p-value for the null.
- The method is more robust (model-free) than t-test (parametric) and more efficient than Wilcoxon test (non-parametric).

- Disadvantage:
  - The resampling nature of the method makes it slow.
  - The tail distribution is difficult to obtain for small replications. e.g. If 2 samples in good prognosis and 3 samples in poor prognosis, there're totally $5!/(2!\times3!)=10$ permutations. The p-value has no enough precision.

Permutation test: simple example

X Y
(78, 72) (102, 105)

$$T = \frac{75 - 103.5}{\sqrt{\frac{18}{2} + \frac{4.5}{2}}} = -8.50$$

What's the null distribution of T?

If X and Y have the same distribution, then T should have the same probability of all the possible permutations.

(78, 72)(102, 105) $\Rightarrow$ T= ?
(78, 102)(72, 105) $\Rightarrow$ T= ?
(78, 105)(72, 102) $\Rightarrow$ T= ?
(72, 102)(78, 105) $\Rightarrow$ T= ?
(72, 105)(78, 102) $\Rightarrow$ T= ?
(102, 105)(78, 72) $\Rightarrow$ T= ?

So p-value of the observed data is 2/6=0.33. Not significant

Permutation test in R

- Package perm (https://cran.r-project.org/web/packages/perm/)
- RMD_example 4.6