Lecture 8: Statistical decision making: hypothesis testing

IST5573

統計方法 Statistical methods

2016/11/2

統計推論 (Statistical inference)

- 統計推論：根據樣本(sample)得到的証據，對母體(population)所要探討性狀或問題作下結論。
- 樣本不是母體，所以根據樣本所的結論帶有不確定性(uncertainty)。
- 不確定性可以抽樣誤差(sampling error)及機率(probability)加以量化。

Two methods of inferences

- Estimation
  1. point estimation: an estimator of a population parameter. E.g., $\bar{x} \rightarrow \mu$.
  2. interval estimation: a point estimate plus an interval that expresses the uncertainty and variability associated with the estimate. E.g., confidence interval.
- Hypothesis testing: given the observed data, do we reject or accept a pre-specified null hypothesis in favor of an alternative?

Gene expression dataset: samplexprs.csv

- van't Veer et al. (2002) (http://www.nature.com/nature/journal/v415/n6871/full/415530a.html)
- 78 breast cancer patients, where 44 patients in good prognosis group and 34 in poor prognosis group

id	age	metastases	followup	ERp	J00129	Contig29982_RC	Contig42854	Contig42
FG80	52	0	7.35	100	-0.795	-0.387	0.199	-0.2
SF58	50	1	1.15	0	-0.509	0.459	-0.257	-0.0
DE72	54	0	12.12	100	-0.961	-0.631	0.037	-0.1
DE65	40	0	6.25	0	-0.749	0.699	-0.346	0.0
HG87	53	0	5.18	0	-0.426	-0.406	-0.355	0.4
HG88	37	1	1.09	100	-0.566	-0.596	-0.352	-0.3

Variable	Description
id	An unique identification number
age	Age at diagnosis of breast cancer (year)
metastases	0=good prognosis group, 1=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

Hypothesis testing for continuous random variables

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 8.2)

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 and poor prognosis groups is equal to 0.
- *Ho* : $\mu_G = \mu_P$ *Ha* : $\mu_G \neq \mu_P$ (RMD_example 8.3)

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 is equal to the variance for the poor prognosis group.
- *Ho* : $\sigma_G = \sigma_P$ *Ha* : $\sigma_G \neq \sigma_P$ (RMD_example 8.3)

ANOVA

- Compare the means among more than two samples, where all samples are independent
- 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 8.4)

Paired t-test

- Compare the difference between two responses measured on the same statistical unit to a given number
- 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 8.5)

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

Measures of association and hypothesis testing for categorical data

Binomial test

- Compare the population proportion to a specified number.
- Perform a test for testing whether or not the population proportion of ER negative (p) is equal to 0.4.

	ERs		
	+ (1)	- (0)	total
number	56	22	78
prop.	0.72	0.28	1

$$\text{ERs} = \begin{cases} 1, & \text{ERp} > 10 \\ 0, & \text{ERp} \leq 10 \end{cases}$$

- $Ho : p = 0.4 \quad Ha : p \neq 0.4$

(RMD_example 8.6)

Prospective study (cohort study 世代研究)

	metastases (outcome)		
ERs (risk)	poor (1)	good (0)	total
+ (1)	21	35	56
- (0)	13	9	22
total	34	44	78

$$\text{ERs}=\begin{cases}1, & \text{ERp} > 10\\ 0, & \text{ERp} \le 10\end{cases}$$

- Start with
  - 56 ERs postive patients
  - 22 ERs negative patients
- After a period of time, identify the numbers of patients who are poor or good groups.

Question: Does ER positive reduce the likelihood of poor prognosis?

- Poor prognosis rates

ERs	
+	21/56 = 0.375
-	13/22 = 0.591
Total	34/78 = 0.436

- calculate a risk ratio or "relative risk"

$$\mathrm{RR} = \frac{\Pr(\mathrm{poor}|\mathrm{ERs}+)}{\Pr(\mathrm{poor}|\mathrm{ERs}-)} = \frac{p_1}{p_2}$$

  p_1 can be estimated by 21/56

  p_2 can be estimated by 13/22

  estimate of $\mathrm{RR} = \widehat{\mathrm{RR}} = \frac{21/56}{13/22} = 0.635$

- 37 percent reduction in poor prognosis!

- $$\begin{cases} \text{RR} = 1 \rightarrow \text{no association} \\ \text{RR} > 1 \rightarrow \text{positive association} \\ \text{RR} < 1 \rightarrow \text{negative association} \end{cases}$$
- hypothesis testing:

$$\begin{cases} \text{Ho: RR} = 1 \\ \text{Ha: RR} \neq 1 \end{cases} \rightarrow \begin{cases} \text{Ho: } p_1 = p_2 \\ \text{Ha: } p_1 \neq p_2 \end{cases}$$

- RMD_example 8.7

x^2 test / Fisher's exact test

	metastases		
ERs	poor (1)	good (0)	total
+ (1)	21 (*a*)	35 (*b*)	56
- (0)	13 (*c*)	9 (*d*)	22
total	34	44	78

- Whether "metastases" is independent of "ERs", test the association between "metastases" and "ERs"
  1. x^2 test if $a, b, c, d \geq 5$
  2. Fisher's exact test if any $a, b, c, d < 5$

- Perform a x^2 test for testing whether or not the difference of the population proportions of being poor prognosis between ERs+ (p_1) and ERs- (p_2) is equal to 0.
- Ho : $p_1 = p_2$ Ha : $p_1 \neq p_2$
- If p < 0.05 (significant), the risk of poor prognosis for patients with ER positive as compared to patients with ER negative is 0.635. In other word, there is a possible 37% reduction in being poor prognosis when ER positive.
- RMD_example 8.8

Retrospective study (case-control study 病例對照研究)

	metastases		
ERs	poor (1)	good (0)	total
+ (1)	21 (*a*)	35 (*b*)	56
- (0)	13 (*c*)	9 (*d*)	22
total	34	44	78

- If, in fact, start with
  - 34 cases (poor prognosis)
  - 44 controls (good prognosis)
- Then, see how many cases with ER positive and how many controls with ER positive

- In case-control study, we cannot estimate $\Pr(\text{poor}|\text{ERs}+)$, therefore, we cannot estimate RR.
- The odds of poor prognosis for ER positive is $\frac{p_1}{1-p_1}$.

  The odds of poor prognosis for ER negative is $\frac{p_2}{1-p_2}$.

  The odds ratio = OR = $\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$.
- The odds ratio can be estimated by

$$\widehat{\text{OR}} = \frac{ad}{bc} = \frac{21 \times 9}{35 \times 13} = 0.415$$

- The estimate of OR is good for both cohort and case-control study.
- When the prevalence $\Pr(\text{poor})$ is small, the odds ratio is approximately equal to the relative risk.

- $$\begin{cases} \text{OR} = 1 \rightarrow \text{no association} \\ \text{OR} > 1 \rightarrow \text{positive association} \\ \text{OR} < 1 \rightarrow \text{negative association} \end{cases}$$
- hypothesis testing:

$$\begin{cases} \text{Ho: OR} = 1 \\ \text{Ha: OR} \neq 1 \end{cases} \rightarrow \begin{cases} \text{Ho: } p_1 = p_2 \\ \text{Ha: } p_1 \neq p_2 \end{cases}$$

- RMD_example 8.7
- Use x^2 test / Fisher's exact test for hypothesis testing
- If p < 0.05 (significant), there is a possible 60% reduction in the odds of being poor prognosis when ER positive.

Matched-pair study

- Samples are not independent.
- Matched pairs (e.g., case-control pair matched on age).

	poor	good
ERs	0	1
	1	1
	0	0
	1	0
	1	1
	1	1
	0	1
	0	1
	0	1
	0	1
	1	0
	1	1
	0	1
	1	1
	1	1
	1	0
	1	1
	1	0
	1	1
	1	1
	0	1
	1	1
	1	1
	0	0
	0	1
	1	1
	1	1
	1	1
	1	1
	0	1
	0	1
	0	1
	1	1

Example: 34 poor prognosis (cases) match the first 34 good prognosis (controls) on age

- The data are displayed in a different type of table

		controls (good) ERs+	controls (good) ERs-	total
cases (poor)	ERs+	17 (*a*)	4 (*b*)	21
	ERs-	11 (*c*)	2 (*d*)	13
	total	28	6	

a, *d*: concordant pairs = same exposure

b, *c*: discordant pairs = different exposure

- The concordant pairs give us no information about differences. We focus on the discordant pairs.
- The estimated odds ratio of being poor prognosis for ERs+ versus ERs- is

$$\widehat{\mathrm{OR}} = \frac{b}{c} = \frac{4}{11} = 0.363$$

McNemar's test

		controls (good)		
		ERs+	ERs-	total
cases	ERs+	17 (*a*)	4 (*b*)	21
(poor)	ERs-	11 (*c*)	2 (*d*)	13
	total	28	6	

- "cases" and "controls" are not independent--use McNemar's test to test the association
- Matched case-control study, and "paired" study

- Perform a McNemar's test for testing if there is association between prognosis and ER status?

$$\begin{cases} \text{Ho: no association} \\ \text{Ha: have association} \end{cases}$$

- If $p < 0.05$ (significant), conclude that there appears to be reduced risk of being poor prognosis for ER positive.
- RMD_example 8.9

Measure of agreement

- Example:
  1. 2 physicians diagnose the same patients. Do physicians agree on diagnosis?
  2. Expression of J00129 and Contig29982_RC on the same patient. Do expressions of two genes agree?
- Expression agreement:

		Contig29982_RC > -0.5		
		yes	no	total
J00129 > -0.5	yes	21	6	27
	no	30	21	51
	total	51	27	78

Question: Is there agreement? How much?

1. Hypothesis test:

   Ho: no agreement between J00129 and Contig29982_RC

   → use McNemar's test

2. The proportion of agreement = $\frac{21+21}{78} = 53.8\%$

   disadvantage:

   - very strongly influenced by the distribution of positive and negative
   - it's possible that there will be a high agreement by chance alone

3. Kappa coefficient: measure of agreement excluding by chance alone

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

where p_0 is the observed proportion of agreement, and p_e is the proportion expected by chance.

- R can calculate κ
- some guidelines:

$$\begin{cases} 0.8 \leq \kappa: \text{almost perfect agreement} \\ 0.6 \leq \kappa < 0.8: \text{substantial agreement} \\ 0.4 \leq \kappa < 0.6: \text{moderate agreement} \\ 0.2 \leq \kappa < 0.4: \text{fair agreement} \\ 0 \leq \kappa < 0.2: \text{slight agreement} \\ \kappa < 0: \text{agreement is same as random} \end{cases}$$

Cohen's kappa test

- $\text{Ho: } \kappa = 0 \qquad \text{Ha: } \kappa \neq 0$
- If not significant, the extent of agreement is same as random.
- RMD_example 8.10

age $\leq$ 40

metastases

ERs		poor	good	total
	+	10	5	15
	-	4	2	6
	total	14	7	21

age $>$ 40

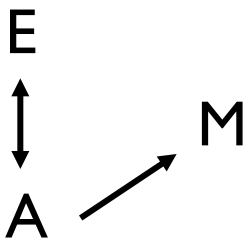
metastases

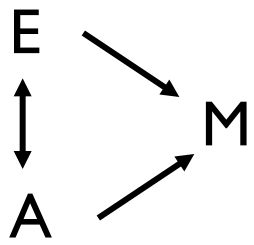
ERs		poor	good	total
	+	11	30	41
	-	9	7	16
	total	20	37	57

When there are more than one table

Confounding

POTENTIAL CONFOUNDER:





causal

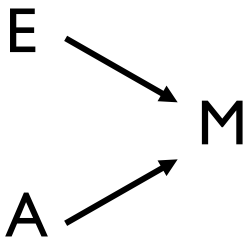
associated

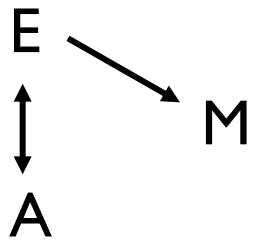
E=ERs (risk)

M=metastases (outcome)

A=age (confounder)

NOT A POTENTIAL CONFOUNDER:





Mantel-Haenszel test

- If the population is stratified (by "age"), we then use Mantel-Haenszel test to test the association between "metastases" and "ERs".
- Under Mantel-Haenszel test, we assume that the odds ratio between "metastases" and "ERs" for "age ≤ 40" is the same as the odds ratio for "age > 40".

Mantel-Haenszel test (cont'd)

- We can use Breslow-Day test for homogeneity of the odds ratios.
- Use Mantel-Haenszel odds ratio (relative risk) to estimate the common odds ratio (relative risk).
- RMD_example 8.11

Nonparametric statistical methods

Nonparametric statistical methods

- A family of probability distributions that can be described by a few parameters is a parametric family.
- Parametric statistical procedures can be used when the sampling distribution is normal or approximately normal.
- A family of probability distributions is nonparametric if it cannot be easily described by a few parameters.

Use of nonparametric methods

- Nonparametric statistical procedures can be used when:
  - sample size(s) are small
  - assumptions of parametric testing procedures cannot be met
- Pro and con of nonparametric methods:
  - pro: insensitive to weird observations (outliers)
  - con: only looking at sign and rank → lose information, less powerful.

Nonparametric versus Parametric

Type of test	Nonparametric	Parametric
one sample test	sign test	one-sample t-test
paired data	Wilcoxon signed-rank test	paired t-test
two sample test	Wilcoxon rank-sum test (Mann-Whitney test)	two-sample t-test

- RMD_example 8.12

Sign test

- Use for one-sample test
- Test 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 : $\mu_{J00129} \neq -0.5$
- It depends on the sign of the differences between observations and given number; not on their actual magnitude

Wilcoxon signed-rank test

- Use for paired data
- Test 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 : $\mu_{J00129} \neq \mu_{Contig29982\_RC}$
- Replaces the observed paired differences with ranks

Wilcoxon rank-sum test

- Related to the Mann-Whitney test
- Used with two independent samples
- Test whether or not the difference of population mean $\log_{10}$ expression intensity ratios on gene J00129 between good and poor prognosis groups is equal to 0.
- Ho : $\mu_G = \mu_P$ Ha : $\mu_G \neq \mu_P$
- The two samples are combined and observations are ranked.

Kruskal-Wallis test

- More than 2 samples
- Test 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

Summary

- Nonparametric statistical methods provide an alternative to parametric methods when the parametric assumptions cannot be met.
- The use of a nonparametric method does not require knowledge of the underlying population distribution(s) or the Central Limit Theorem.
- A nonparametric test result is usually more conservative than a parametric test result.
- Nonparametric methods are often used with laboratory studies and small sample sizes.

Computational methods

Permutation test

- Permutation method can be an extension from any powerful testing methods.
- The null distribution is calculated by randomly permuting the class labels.
- We can the estimate the p-value of a test.
- The method is more robust (model-free) than t-test and more efficient than Wilcoxon test.

- 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

- Details in R:

  Permutation Test – Rstudio (https://www.youtube.com/watch?v=nq3zC4dt6gc)
- And many others

Bootstrap

- Bootstrap methods are a class of nonparametric Monte Carlo methods that estimate the distribution of a population by resampling.
- Resampling methods treat an observed sample as a finite population, and random samples are generated (resampled) from it to estimate population characteristics and make inferences about the sampled population.
- Bootstrap methods are often used when the distribution of the target population is not specified; the sample is the only information available.

Bootstrap

- The bootstrap can estimate the standard error of the estimator of interest.
- Details in R:

  Quick-R: Bootstrapping (http://www.statmethods.net/advstats/bootstrapping.html)
- And many others