Spring 2014
Stat 8004: Statistical Methods II
Lecture 7

1 Review of Last Lecture

Contingency Table:

	Response		
Exposure	+	−	
1	a	b	n_1
2	c	d	n_2
	m_1	m_2	N

Commonly Used Distributions:

- No margin fixed: Poisson
- One margin fixed: Multinomial
- Two margin fixed: Product binomial
- Three margin fixed: Conditional hypergeometric

Measure of Risks:

- Risk difference (RD), usually used in the prospective design
- Relative risk (RR), usually used in the prospective design
- Odds ratio (OR), used in prospective or retrospective design

Inference:

- Large sample, product binomial model
  - Hypothesis testing and confidence limits for RD, RR and OR.

  - Under the null, we can use better estimators of variance than under the alternative.
  - For RR and OR, we do log-transformation before inference and then transform them back. This is to make the range of the confidence limits bounded.

- Small sample (Exact Inference), conditional hypergeometric model

2 Inference

2.1 Other Large Sample Hypothesis Testing Methods

2.1.1 Pearson Contingency Chi-Square Test

Pearson Chi-square test is one of the most commonly used test for $R \times C$ contingency table.

$$X_P^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}},$$

where O_{ij} is the observed frequency in the ith row and jth column and $\hat{E}_{ij}$ is the estimated expected frequency under the null hypothesis.

We would like to know whether the exposure is associated with the outcome. How to set up the hypothesis?

$$\mathrm{H}_0 : \ \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}.$$

Under the null,

$$E_{ij} = \mathbb{E}\left(O_{ij}\right) = n\pi_{ij} = n\pi_{i\cdot}\pi_{\cdot j}.$$

Substituting the sample estimators: $\hat{\pi}_{i\cdot} = n_{i\cdot}/n$ and $\hat{\pi}_{\cdot j} = n_{\cdot j}/n$ yields the estimated expected frequencies:

$$\hat{E}_{ij} = n_{i\cdot}n_{\cdot j}/n.$$

Under the null, asymptotically $X_P^2 \xrightarrow{d.} \chi^2_{(R-1)(C-1)}$.

For a 2×2 table, $|O_{ij} - \hat{E}_{ij}|$ is constant for all celles of the table. Thus,

$$X_P^2 = \frac{(ad - bc)^2 N}{n_1 n_2 m_1 m_2},$$

which is asymptotically distributed as $\chi^2(1)$ under H_0.

Example: Smoking and Lung Cancer. $X_P^2 = 46.99$, $df = 1$, the $p - value = 7.14E - 12$. We reject H_0 under the level 0.05.

2.2 Conditional Mantel-Haenszel Test

Consider the Hypergeometric distribution. We discussed the exact test just now. We can also build up large sample test under this model.

Under $\mathrm{H}_0: \ \varphi = OR = 1$,

$$\begin{aligned}\mathbb{E}(a) &= \frac{n_1 m_1}{n} \\ \widehat{\mathrm{Var}}(a) &= \frac{n_1 n_2 m_1 m_2}{n^2(n-1)}\end{aligned}$$

Conditional Mantel-Haeszel Test:

$$X_c^2 = \frac{(a - \mathbb{E}(a))^2}{\widehat{\mathrm{Var}}(a)},$$

and the corresponding z-statistics is

$$Z_c = \frac{a - \mathbb{E}(a)}{\sqrt{\widehat{\mathrm{Var}}(a)}}.$$

Under H_0, $X_c^2 \xrightarrow{d} \chi^2(1)$ and $Z_c \xrightarrow{d} N(0,1)$.

Example: Smoking and Lung Cancer. $Z_c = 6.93 > z_{0.975}$, and therefore, we reject H_0.

2.3 Cochran's Test

Consider the Product Binomial Model. Here is another option to test $\mathrm{H}_0: \ \pi_1 = \pi_2 = \pi$. Instead of testing on the risks, we develop tests based on $[a - \mathbb{E}(a)]$.

Under the Product Binomial Distribution,

$$\begin{aligned}\mathbb{E}(a) &= n_1 \pi_1 \\ \mathrm{Var}(a) &= n_1 \pi_1 (1 - \pi_1).\end{aligned}$$

They can be reasonably estimated as

$$\begin{aligned}\widehat{\mathbb{E}}(a) &= n_1 \hat{\pi}_1 = n_1 m_1 / n \\ \widehat{\mathrm{Var}}(a) &= n_1 m_1 m_2 / n^2\end{aligned}$$

Likewise,

$$\begin{aligned}\mathrm{Var}(c) &= n_1 \pi_2 (1 - \pi_2). \\ \widehat{\mathrm{Var}}(c) &= n_2 m_1 m_2 / n^2\end{aligned}$$

Based on these results, under $H_0: \pi = \pi_1 = \pi_2$.

$$V_u = \text{Var}[a - \hat{E}(a)] = \frac{n_2^2 \text{Var}(a) + n_1^2 \text{Var}(c)}{n^2} = \frac{n_1 n_2 \pi(1-\pi)}{n},$$

which can be estimated as

$$\hat{V}_u = \widehat{\text{Var}}[a - \widehat{\mathbb{E}}(a)] = \frac{n_1 n_2 \hat{\pi}(1-\hat{\pi})}{n} = \frac{n_1 n_2 m_1 m_2}{n^3}.$$

The Cochran's Chi-square test statistic for 2×2 table is

$$X_u^2 = \frac{[a - \widehat{\mathbb{E}}(a)]^2}{\hat{V}_u},$$

which can be shown to be equal to X_P^2. The corresponding z-statistic is

$$Z_u = \frac{a - \widehat{\mathbb{E}}(a)}{\sqrt{\hat{V}_u}}.$$

Under H_0, asymptotically X_u^2 follows $\chi^2(1)$ and Z_u follows $N(0,1)$.

2.4 Likelihood Ratio Test

Like the Pearson test, this test tests the independence of the row and the column factors. It can be adopted to $R \times C$ contingency table.

$$G^2 = 2 \sum_{i=1}^{R} \sum_{j=1}^{C} O_{ij} \log \left(\frac{O_{ij}}{\hat{E}_{ij}} \right),$$

where O_{ij} and $\hat{E}_{ij}$ are observed and estimated expected frequencies defined above. Asymptotically, under $H_0: \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$, $G^2 \sim \chi^2((R-1)(C-1))$.

Example: Smoking and Lung Cancer. $G^2 = 46.73 > z_{\chi^2(1),0.95}$. And therefore, we reject H_0.

2.5 SAS Code

Example: Smoking and Lung Cancer

```
data LungCancer;
input group $ resp $ n;
cards;
```

		Lung Cancer?		
		Yes	No	Total
Smoking?	Yes	360	120	480
	No	95	105	200
	Total	455	225	680

Table 1: Contingency table for smoking and lung cancer data

```
smoking cancer 360
non-smoking cancer 95
smoking non-cancer 120
non-smoking non-cancer 105;
run;

proc freq data = LungCancer;
tables group resp / chisq cl cmh relrisk riskdiff expected;
weight n;
run;
```

3 Diagnostic Tests

3.1 Sensitivity and Specificity

Review of Conditional Probability:

$$\begin{aligned}\mathbb{P}(A \mid B) = &\text{ probability that a randomly selected individual} \\ &\text{ has property A, given that he/she has property B} \\ &= \text{read as ``probability A exists given that B exists''} \\ &= \frac{\mathbb{P}(A \text{ and } B)}{\mathbb{P}(B)}, \text{ provided } \mathbb{P}(B) \neq 0\end{aligned}$$

- Association $\Rightarrow$ when a person has one characteristic (B), chances of having the other (A) are affected; $\mathbb{P}(A \mid B) \neq \mathbb{P}(A)$
- Independence $\Rightarrow$ having B does not affect the chance of having A; $\mathbb{P}(A \mid B) = \mathbb{P}(A)$
- Law of Total Probability: for event A and strata B_k, $k = 1, \ldots, K$, *s.t.* $\sum_{k=1}^{K} \mathbb{P}(B_k) = 1$,
$$\mathbb{P}(A) = \mathbb{P}(A \mid B_1)\mathbb{P}(B_1) + \ldots + \mathbb{P}(A \mid B_K)\mathbb{P}(B_K)$$
In particular, $\mathbb{P}(A) = \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid \bar{B})\mathbb{P}(\bar{B})$.
- Bayes Theorem: for events A and B,
$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

Screening test is a commonly seen test in biomedical studies. Patients or subjects receive screening tests to pre-diagnose whether she/he has the disease of interest. Suppose

$$\begin{aligned} D &= \text{person has disease in question;} \\ N &= \text{person does not have disease;} \\ T+ &= \text{person gives positive test response;} \\ T- &= \text{person gives negative test response;} \end{aligned}$$

What are $\mathbb{P}(T+ \mid D)$ and $\mathbb{P}(T- \mid N)$?

Due to Yerushalmy (1947)

	Disease Present (D)	Disease absent (N)
Test Positive ($T+$)	True Positive	False Positive
Test Negative ($T-$)	False Negative	True Negative

Therefore, $\mathbb{P}(T+ \mid D)$ is called *sensitivity* since a larger value implies the test is more sensitive; and $\mathbb{P}(T- \mid N)$ is called *specificity*, since a larger value implies the test is more specific. Sensitivity is also called *true positive rate* (TPR), and 1-specificity is also called *false positive rate* (FPR).

Usually, before the test is conducted, we can specify a *pre-test probability* of disease. It could be an estimated belief of having disease BEFORE the test is performed. For example, it could be the prevalence of disease in population. From a Bayesian point of view, it could be a prior probability of having disease.

As stated above, sensitivity and specificity tell how well the test discriminates patients with and without disease.

Now, given the prior, sensitivity and specificity, what is our belief that the patient has the disease?

Post-test Probability:

- Positive Predictive Value (PPV):

$$PPV = \mathbb{P}(D \mid T+) = \frac{\mathbb{P}(T+ \mid D)\,\mathbb{P}(D)}{\mathbb{P}(T+)}$$

  This is the post-test probability that the patient has the disease.

- Negative Predictive Value (NPV):

$$NPV = \mathbb{P}(N \mid T-) = \frac{\mathbb{P}(T- \mid N)\,\mathbb{P}(N)}{\mathbb{P}(T-)}$$

  This measures how well the test rules out the disease.

Both PPV and NPV are functions of pre-test probability, sensitivity and specificity. Let

$$\Pi = \text{pre-test probability} = \text{ prevalence of disease in population.}$$

Then,

$$PPV = \frac{\text{Sens} \cdot \Pi}{\text{Sens} \cdot \Pi + (1 - \text{Sens}) \cdot (1 - \Pi)}$$
$$NPV = \frac{\text{Spec}(1 - \Pi)}{\text{Spec} \cdot (1 - \Pi) + (1 - \text{Sens}) \cdot \Pi},$$

Hypothetical Example: Sampled 1000 people known to have the disease, and 1000 known to not have the disease:

	Disease Status		
Test Result	Present (D)	Absent (N)	Total
$T+$	950	10	960
$T-$	50	990	1040
Total	1000	1000	2000

- Sensitivity $= \mathbb{P}(T+ \mid D) = 950/1000 = 0.95$.
- Specificity $= \mathbb{P}(T- \mid N) = 990/1000 = 0.99$.
- PPV $= \frac{0.95\Pi}{0.95\Pi + 0.01(1-\Pi)} = \frac{95\Pi}{94\Pi+1}$.
- NPV $= \frac{0.99(1-\Pi)}{0.99(1-\Pi)+0.05\Pi} = \frac{99-99\Pi}{99-94\Pi}$.
- Values for PPV and NPV vary based on prevalence (number of people having disease in population)

Suppose the sensitivity and specificity is unchanged. How would the prevalence of disease in population change PPV and NPV?

Π	PPV	NPV
1/1,000,000	0.0001	1.0
1/100,000	0.0009	1.0
1/10,000	0.0094	0.99999
1/1,000	0.087	0.99995
1/500	0.160	0.99990
1/200	0.323	0.99975
1/100	0.490	0.99949

- If disease is not too prevalent, NPV very high but PPV rather small
- What do you think of the test?

Sometime, the test have multiple cut-offs. How to show sensitivity and specificity for different cut-offs?

Example: Hypothyroidism. Goldstein and Mushlin (JGIM, 1987) measured thyroxine (T4) values in patients suspected of hypothyroidism.

T4 value	Hypothyroid	Euthyroid
$[0, 5]$	18	1
$(5, 7]$	7	17
$(7, 9]$	4	36
$(9, \infty)$	3	39
Total	32	93

How do sensitivity and specificity change depending on choice of T4 level (cut-off defining hypothyroidism)?

Let's consider various cutoffs.

- For cutoff = 5,

Cut-off	D	ND	Total
≤ 5 (T+)	18	1	19
> 5 (T−)	14	92	106
Total	32	93	125

, Sens = 0.56, Spec = 0.99.

- For cutoff = 7,

Cut-off	D	ND	Total
≤ 7 (T+)	25	18	43
> 7(T−)	7	75	82
Total	32	93	125

, Sens = 0.78, Spec = 0.81.

- For cutoff = 9,

Cut-off	D	ND	Total
≤ 9 (T+)	29	54	83
> 9 (T−)	3	39	42
Total	32	93	125

, Sens = 0.91, Spec = 0.42.

For different cut-offs, as sensitivity increases, specificity decreases; and vise-versa. Plot sensitivity (TPR) vs. (1-specificity) (FPR) can help visualize the results. Such plot is called *receiver operating characteristic* (ROC) curve. Figure 1 is an example. ROC curve is

- Concave, connecting $(0, 0)$ and $(1, 1)$
- summarizes predictive power for all possible cutoffs
- shows tradeoff between sensitivity and specificity
- tangent line slope at cutoff equals likelihood ratio for that value of test (will discuss LR later)

Please refer to `ROC.pdf` for the ROC curve for the T4 cell example.

To compare ROC curves of different tests. We need to consider the *area under a ROC curves* (AUC). The larger the AUC, the better the test/prediction. We can view AUC as concordance index or discrimination, *i.e.*, probability of randomly drawn patient pairs (one with disease and one without) for which test correctly classifies each. AUC is bounded by

$$0.5 \leq \text{AUC} \leq 1.$$

If AUC equal to 0.5, then the test/prediction is no better than random guessing; if AUC = 1, the test is perfect.

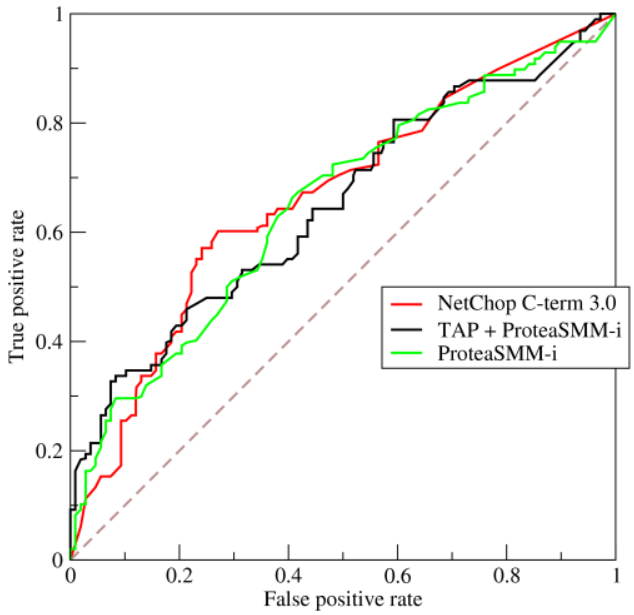

Figure 1: ROC curve

3.2 Likelihood Ratios

3.2.1 Definition

We discuss an alternative method (to sensivity and specificity) of assessing performance of diagnostic test. It can handle multiple-outcome tests. The measure we introduce here is *likelihood ratio* (LR),

$$LR = \frac{\mathbb{P}\left(\text{test outcome} \mid \text{patient has disease}\right)}{\mathbb{P}\left(\text{test outcome} \mid \text{patient is disease-free}\right)}$$

To be more specific, Likelihood ratio of a positive test:

$$LR^{+} = \frac{\mathbb{P}\left(T+ \mid D\right)}{\mathbb{P}\left(T+ \mid N\right)} = \frac{\text{Sens}}{1 - \text{Spec}}$$

And, Likelihood ratio of a negative test:

$$LR^{-} = \frac{\mathbb{P}\left(T- \mid D\right)}{\mathbb{P}\left(T- \mid N\right)} = \frac{1 - \text{Sens}}{\text{Spec}}$$

Note that LR^{+} is the larger the better and LR^{-} is the smaller the better.

Likelihood ratio can simplify the caculation of post-test probability. Recall that

$$Odds = Probability/(1 - Probability),$$

and

$$Probability = Odds/(1 + Odds).$$

The post-test odds are

$$\frac{\mathbb{P}(D \mid T+)}{\mathbb{P}(N \mid T+)} = \frac{\mathbb{P}(T+ \mid D)\,\mathbb{P}(D)}{\mathbb{P}(T+)} \cdot \frac{\mathbb{P}(T+)}{\mathbb{P}(T+ \mid N)\,\mathbb{P}(N)} = \frac{\mathbb{P}(D)}{\mathbb{P}(N)} \cdot \frac{\mathbb{P}(T+ \mid D)}{\mathbb{P}(T+ \mid N)}.$$
$$\frac{\mathbb{P}(D \mid T-)}{\mathbb{P}(N \mid T-)} = \frac{\mathbb{P}(T- \mid D)\,\mathbb{P}(D)}{\mathbb{P}(T-)} \cdot \frac{\mathbb{P}(T-)}{\mathbb{P}(T- \mid N)\,\mathbb{P}(N)} = \frac{\mathbb{P}(D)}{\mathbb{P}(N)} \cdot \frac{\mathbb{P}(T- \mid D)}{\mathbb{P}(T- \mid N)}.$$

Thus,

$$\text{post-positive-test odds} = \text{pre-test odds} \cdot LR^+$$
$$\text{post-negative-test odds} = \text{pre-test odds} \cdot LR^-$$

And also,

$$PPV = \frac{\text{post-positive-test odds}}{1 + \text{post-positive-test odds}}, \quad NPV = \frac{1}{1 + \text{post-negative-test odds}}.$$

Example: The following are results fo a systematic review of serum ferritin as diagnostic test for iron deficiency anaemia. Suppose the data is randomization data, and the prevalence can be estimated as

$$\text{Prevalence} = 809/2579 = 0.31.$$

Please note that the prevalence can NOT be estimated if the data is not popultion or randomized data.

	Target Disorder		
Test Result	Present (D)	Absent (N)	Total
$T+$ ($<$ 65mmol/L)	731	270	1001
$T-$ ($\geq$ 65 mmol/L)	78	1500	1578
Total	809	1770	2579

- $\Pi = 0.31$
- Sensitivity $= 731/809 = 0.90$
- Specificity $= 1500/1770 = 0.85$
- PPV $= 0.73$
- NPV $= 0.95$

Now we use the likelihood ratio method.

- Pre-test odds $= 0.31/(1 - 0.31) = 0.45$
- $LR^+ = 0.90/(1 - 0.85) = 6$
- $LR^- = (1 - 0.90)/0.85 = 0.12$
- Post-positive-test odds $= 0.45 \times LR^+ = 2.7$
- Post-negative-test odds $= 0.45 \times LR^- = 0.054$
- $PPV = 2.7/(1 + 2.7) = 0.73$
- $NPV = 1/(1 + 0.95) = 0.95$

Likelihood ratio provides insight on how test result affects disease likelihood:

- LRs > 10 or < 0.1 cause major differences between pre- and post-test probabilities of having disease
- LRs between 5-10 or 0.1-0.2 cause moderate differences
- LRs between 0.5-2 cause minor to no change
- LR $= 1$ implies no change

3.2.2 Series Tests

Likelihood ratios are extremely helpful when there are a series of tests. From the first test,

$$\text{Post-test Odds}_1 = \text{LR}_1 \times \text{Pre-test Odds}$$

From the second test, set Post-test Odds_1 as new "pre-test odds".

$$\text{Post-test Odds}_2 = \text{LR}_2 \times \text{Post-test Odds}_1$$

Continuing this process after n tests:

$$\begin{aligned}\text{Post-test Odds}_n &= \text{LR}_n \times \text{Post-test Odds}_{n-1} \\ &= \text{LR}_n \times \text{LR}_{n-1} \times \text{Post-test Odds}_{n-2} \\ &= \text{LR}_n \times \text{LR}_{n-1} \times \ldots \times \text{LR}_1 \times \text{Pre-test Odds}\end{aligned}$$