Statistics Primer II

David H. Alexander

August 17, 2010

Overview

- Contingency tables (categorical data)
- Methods of evaluating significance
  - Large-sample tests
  - Exact tests
  - Permutation tests
- Trend tests (ordinal data)
- Linear regression (continuous data)

Contingency tables

A contingency table allows us to cross-classify a categorical outcome and (one or more) categorical factors. A 2×2 two-way classification example:

	Myocardial infarction		
	Yes	No	Total
Aspirin	104	10,933	11,037
Placebo	189	10,845	11,034
Total:	293	21,778	22,071

In an $I \times J$ two-way classification table, factor has I *levels*, outcome has J levels. If $J = 2$, a *binary* outcome.

Contingency table cell counts

Cell counts (*frequencies*) denoted $\{n_{ij}\}$.

	Outcome		
	+	-	Total
Treatment	n_{11}	n_{12}	n_{1+}
Control	n_{21}	n_{22}	n_{2+}
Total:	n_{+1}	n_{+2}	n

n_{ij} generated from joint probability model π_{ij} or from a (conditional) *success probability* π_i for each row. Depends on sampling scheme.

Odds ratio

$$OR \doteq \frac{\text{Odds}_{\text{treatment}}}{\text{Odds}_{\text{control}}} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

$$\widehat{OR} \doteq \frac{p_1/(1-p_1)}{p_2/(1-p_2)}, \quad \text{where } p_1 \doteq n_{11}/n_{1+}, \text{ etc.}$$

$$= \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Interpretation for disease studies:

- $OR = 1$: treatment factor and outcome are independent
- $OR < 1$: odds of disease smaller under treatment
- $OR > 1$: odds of disease greater under treatment

Does aspirin help?

Does taking aspirin make it more or less likely you will have a heart attack?

$$\widehat{OR} = \frac{(104)(10,845)}{(10,933)(189)} = 0.546$$

So there is some evidence associating aspirin treatment and decreased odds of a heart attack.

But is there *enough* evidence to make the conclusion? i.e. is the observed relationship significant?

Evaluating significance

Is there a significant association between levels of the factor, and the outcome?

Test the null hypothesis of *independence* between the factor and the outcome.

$$
\begin{aligned}
&H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \\
&H_a : \pi_{ij} \text{ unconstrained (but rows, columns must sum to one)}
\end{aligned}
$$

How can we perform the test?

A large-sample test: Pearson's χ^2

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

where $\{\hat{\mu}_{ij}\}$ are estimates of the expected cell counts under H_0,

$$\hat{\mu}_{ij} \doteq np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n}$$

Asymptotically, follows a χ^2 distribution with $(I-1)(J-1)$ degrees of freedom.

Aspirin example: $X^2 = 25.01, df = 1 \implies p = 5.7e - 07$

The (potential) problem with using a large sample test

The distribution of the test statistic X^2 is only close to the χ^2 distribution for sufficiently large sample size.

Rule of thumb: χ^2 test should only be trusted to yield reliable results (p values) when all $\hat{\mu}_{ij} \geqslant 5$.

Exact tests

Fortunately, when the sample is small, we can readily compute the *exact* p value, not relying on any distributional approximations on X^2.

How: compute the total probability, under H_0, of all configurations that retain the same marginal totals as observed table, while achieving a value of the X^2 statistic larger than that observed.

Fisher's exact test

For the 2×2 case, this is called *Fisher's exact test*. Since totals fixed, n_{11} determines all other entries, and the probability of any configuration n_{11} can be written as

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1} - n_{11}}}{\binom{n}{n_{+1}}},$$

a *hypergeometric* probability.

The sum of the appropriate probabilities is computed by software (R, MENDEL).

Permutation tests

General mechanism for calculating the significance level of observed data. Not limited to contingency tables. Very useful when the distribution of the test statistic is complicated or unknown.

Case/control association example: for each individual i we have measured factors/covariates $\mathbf{x}_i$ and observed outcome y_i.

Randomly permute the outcome vector $\mathbf{y} = \{y_i\}$ and recompute the measure of association (e.g. the χ^2 statistic) as X_*^2.

Repeat this procedure B times (B is large).

$$\text{Permutation p value} = \frac{\#\text{ of times } X_*^2 \geqslant X^2}{B}$$

Permutation test vs. exact test

Both are nonparametric alternatives to large-sample tests.

Exact tests require enumerating *all* configurations that are at least as favorable to H_a as the observed data. Useful when sample size is small.

Permutation tests require enumerating a large number B of configurations. p value is inexact—a *Monte Carlo* estimate of the true p value. Useful in general situations where, for example, the distribution of the test statistic is complicated or unknown.

Ordinal data and trend tests

Sometimes we have a categorical factor whose levels have an implicit ordering, i.e.

- 0 drinks/day $<$ 1-2 drinks/day $<$ 3-5 drinks/day
- dd $<$ Dd $<$ DD

An *ordinal* factor. If we assign a *score* to each factor level, we can perform a *trend test* which has fewer degrees of freedom than the simple χ^2 test of independence and can thus be more powerful.

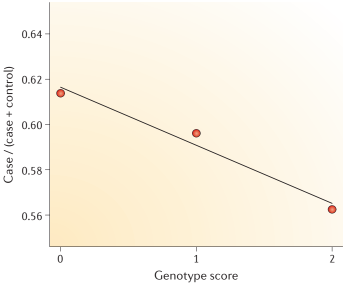
The test statistic is the squared Pearson correlation between the factor score and the outcome.

Cochran-Armitage trend test

When outcome is binary, this is called the *Cochran-Armitage trend test*.

Used in the additive model of genetic penetrance:

$$\mathsf{Score}(Dd) = \frac{1}{2}\left[\mathsf{Score}(dd) + \mathsf{Score}(DD)\right]$$



Testing if slope of line is zero.

For genotypes, $df = 1$, versus $df = 2$ for more general χ^2 test.

Figure credit: Balding reference

Linear regression

Linear regression models a linear relationship between a *continuous* variable X and a *continuous* outcome Y.

For example, could be used to model the relationship between height and weight, or the relationship between blood LDL and HDL lipid levels.

$$y_i = \mu + \beta x_i + \epsilon_i$$

Linear regression inference

Effect size and direction

$$\hat{\beta} \doteq \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

yields the *least squares line* minimizing

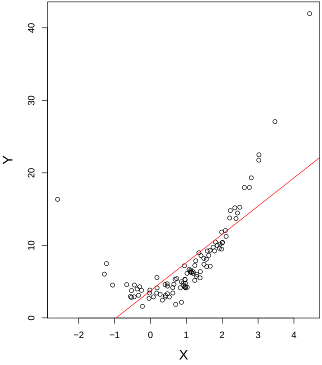
$$SSE = \sum_i e_i^2, \quad \text{where } e_i = y_i - (\hat{\mu} + \hat{\beta} x_i)$$
$$= y_i - \hat{y}_i$$

Significance

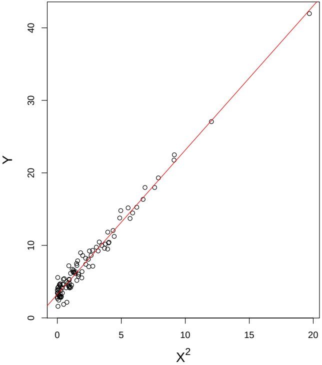
We often want to test a simple hypothesis like $H_0 : \beta = 0$. Use a *t*-test in this case. More complicated tests in the multivariate case use *F*-tests.

Nonlinear relationships and transformations

If the true relationship is nonlinear, can still use linear regression after *transforming* X or Y.



Y vs.X



Y vs. X^2

Transformations often tried: log, powers. Search can be automated (e.g. Box-Cox transform).

Recommended reading

- A. Agresti.
  An introduction to categorical data analysis.
  Wiley-Blackwell, 2007.

- T.A. Pearson and T.A. Manolio.
  How to interpret a genome-wide association study.
  JAMA, 299(11):1335, 2008.

- D.J. Balding.
  A tutorial on statistical methods for population association studies.
  Nature Reviews Genetics, 7(10):781–791, 2006.

- L. Wasserman.
  All of statistics: a concise course in statistical inference.
  Springer Verlag, 2004.