STAT 135, Concepts of Statistics

Helmut Pitters

Hypothesis testing 3

Department of Statistics
University of California, Berkeley

March 14, 2017

Hypothesis testing.

Uniformly most powerful test. Neyman-Pearson lemma gives most powerful test provided both hypotheses are simple. In general not possible to find most powerful test for composite hypotheses. However, if null is simple, can extend theory to so-called uniformly most powerful tests.

Definition (Uniformly most powerful test)

A test

$$H_0\colon \theta = \theta_0 \qquad H_A\colon \theta \in \Theta_A$$

with simple null is called *uniformly most powerful (UMP)* if it is most powerful for every simple alternative in H_A, i.e. if for any $\theta_a \in \Theta_A$ it is most powerful for

$$H_0\colon \theta = \theta_0 \qquad H_A\colon \theta = \theta_a.$$

Hypothesis testing.

Generalized likelihood ratio test. If hypotheses are composite

$$H_0 \colon \theta \in \Theta_0 \qquad H_A \colon \theta \in \Theta_A$$

cannot expect to find most powerful test. However, we can still compare likelihoods of null and alternative, both evaluated at their maximal value. This leads to what is called the generalized likelihood ratio tests which are similarly important to testing as MLEs are in estimation.

Definition (Generalized likelihood ratio test)

The *generalized likelihood ratio test (GLRT)* rejects H_0 for small values of the *generalized likelihood ratio (GLR)*

$$\Lambda := \frac{\sup_{\theta \in \Theta_0} \operatorname{lik}(\theta)}{\sup_{\theta \in \Theta} \operatorname{lik}(\theta)},$$

where $\Theta = \Theta_0 \cup \Theta_A$.

As before, reject H_0 for small values of Λ.

Hypothesis testing.

Generalized likelihood ratio test. To work out the critical value of the GLRT for fixed level α we would need to know the sampling distribution $(\Lambda | H_0)$ of Λ under H_0.
This distribution has no simple form in general, but $-2\log\Lambda$ can be approximated by a chi squared distribution for large sample size.

Fact

Under smoothness conditions on $f(x|\theta)$ the null distribution of

$$-2\log\Lambda$$

is asymptotically (as $n \to \infty$) distributed according to a chi squared distribution with

$$\textit{\#free parameters in } \Theta - \textit{\#free parameters in } \Theta_0$$

degrees of freedom.

Hypothesis testing.

Goodness of fit

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit). Emergency room of large hospital assigns patients to one of three categories:

1. Stable. No immediate treatment required.
2. Serious. Immediate treatment not required, but patient needs to be monitored until physician available.
3. Critical. Patient's life endangered without immediate treatment.

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit). Hospital records over past year show that

- 50% of patients classified stable,
- 30% of patients classified serious,
- 20% of patients classified critical,

hence if we took a random sample of n patients and observed

$$N_i \text{ patients in category } i$$

(where $\sum_i N_i = n$) we expect that

$$(N_1, N_2, N_3)$$

follows a multinomial distribution with parameters $(3, 50\%, 30\%, 20\%)$.

Hypothesis testing.

Review: multinomial distribution. Think of n different marbles that we paint in c different colors. We paint marbles one after the other, with

$$\mathbb{P}\left\{\text{a particular marble is painted in color } i\right\} = p_i$$

(with $\sum_i p_i = 1$, $p_i \geq 0$) independent of the colors in which the other marbles are painted. Let

$$X_i \coloneqq \text{\# marbles of color } i.$$

Then

$$\mathbb{P}\left\{X_1 = x_1, \ldots, X_c = x_c\right\} = \binom{n}{x_1, \ldots, x_n} \prod_{i=1}^{c} p_i^{x_i}$$

if $\sum_i x_i = n$ and $= 0$ otherwise. The vector $(X_1, \ldots, X_n)$ has a *multinomial distribution* with parameters $(n, p_1, \ldots, p_c)$ denoted

$$(X_1, \ldots, X_n) \sim \mathrm{multinomial}(n, p_1, \ldots, p_c).$$

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit). Emergency room's good reputation resulted in increased number of patients.

Important question for organization: has increased number of patients also brought about a change in distribution of patients among categories?
I.e. want to test

$$H_0\colon (p_1, p_2, p_3) = (0.5, 0.3, 0.2) \quad \text{vs. } H_A\colon (p_1, p_2, p_3) \neq (0.5, 0.3, 0.2).$$

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit). Record categories of n incoming patients, modeled as independent random draws from $1, 2, 3$.

$$O_i := \text{\# patients in sample of category } i$$

Under null expect

$$E_i := \mathbb{E}[N_i] = np_i$$

patients in category i.[1] *Pearson's chi-squared test* is based on test statistic

$$\sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i},$$

which can be approximated by a χ^2_{c-1} distribution if sample size n is large enough.[2]

[1] Recall that the marginal distribution X_i in the multinomial distribution $(X_1, \ldots, X_c) \sim \mathrm{multinomial}(c, p_1, \ldots, p_c)$ has a $\mathrm{binomial}(n, p_i)$ distribution.

[2] As a rule of thumb: approximation is good if $O_i \geq 5$.

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit). A random sample of $n = 200$ patients is taken and their observed frequencies are as in the table below

	stable	serious	critical
observed frequencies (O_i)	98	48	54
expected frequencies (E_i)	100	60	40

Table: Frequencies of patients in different categories.

Notice that if $(p_1, p_2, p_3) = (0.5, 0.3, 0.2)$, we expect

$$
\begin{aligned}
E_1 &:= \mathbb{E}[N_1] = 200 \times 0.5 = 100 \text{ patients in category } 1 \\
E_2 &:= \mathbb{E}[N_2] = 200 \times 0.3 = 60 \text{ patients in category } 2 \\
E_3 &:= \mathbb{E}[N_3] = 200 \times 0.2 = 40 \text{ patients in category } 3
\end{aligned}
$$

since $N_i \sim \text{binomial}(200, p_i)$.

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit). Heuristically, if there are large differences

$$|O_i - E_i|$$

between observed and expected numbers of patients for some categories, we reject the null. However, if these differences are small, the data do not provide sufficient evidence to reject H_0. Pearson's chi squared test is based on the test statistic

$$X^2 := \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i},$$

which is well approximated by a χ^2_{c-1} distribution provided the sample size n is large. Null is rejected for large values of X^2.

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit).
Define the α-percentile of χ_n^2 by

$$\mathbb{P}\left\{Y \leq \chi_n^2(\alpha)\right\} = \alpha$$

for $Y \sim \chi_n^2$.
Fix significance level $\alpha = 0.05$. We reject H_0 if

$$X^2 > \chi_{c-1}^2(1-\alpha) = \chi_2^2(0.95) = 5.99.^{3}$$

In our example

$$X^2 = \frac{(98-100)^2}{100} + \frac{(48-60)^2}{60} + \frac{(54-40)^2}{40} = 7.34,$$

thus we reject H_0 on the basis of the data, i.e. there is strong evidence that the distribution of patients among categories has changed.

[3] Find percentiles of chi squared distribution tabulated or evaluate them via statistical software package.

Hypothesis testing.

Pearson's chi squared test that we saw in the last example can be applied to a broad range of settings.

Many observations in social and physical sciences are not numerical, but can be assigned to different categories. This is called categorial or enumerative data. E.g.

- brand of motor vehicle on certain highway section
- classification of documents according to topics
- classification of animals according to species
- blood type of a person: A, B, AB, O

Hypothesis testing.

Pearson's chi squared test can be readily generalized to some arbitrary number c of categories.

Consider population with each individual belonging to one of c categories
(e.g. residents of California with blood types A, B, AB, O), and let

$$p_i := \text{\# relative frequency of items in category } i.$$

Draw random sample of size n.

$$N_i := \text{\# items of category } i \text{ in sample.}$$

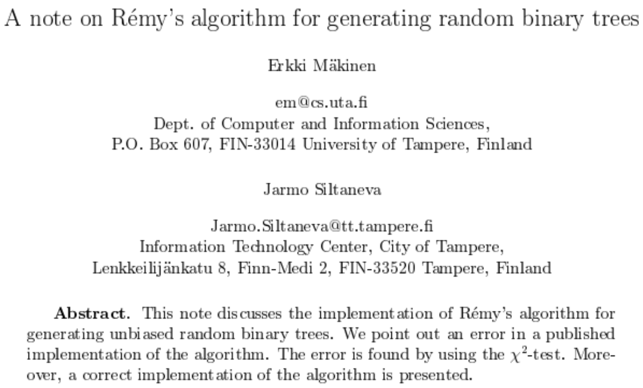
$$(N_1, \ldots, N_c) \sim \mathrm{multinomial}(n, p_1, \ldots, p_c).$$

The hypotheses

$H_0\colon (p_1, \ldots, p_c) = (\pi_1, \ldots, \pi_c) \qquad H_A\colon (p_1, \ldots, p_c) \neq (\pi_1, \ldots, \pi_c)$

(for some $\pi_i \geq 0$ s.t. $\sum_i \pi_i = 1$) can then be tested via Pearson's chi squared test.

Hypothesis testing.



A note on Rémy's algorithm for generating random binary trees

Erkki Mäkinen

em@cs.uta.fi
Dept. of Computer and Information Sciences,
P.O. Box 607, FIN-33014 University of Tampere, Finland

Jarmo Siltaneva

Jarmo.Siltaneva@tt.tampere.fi
Information Technology Center, City of Tampere,
Lenkkeilijänkatu 8, Finn-Medi 2, FIN-33520 Tampere, Finland

Abstract. This note discusses the implementation of Rémy's algorithm for generating unbiased random binary trees. We point out an error in a published implementation of the algorithm. The error is found by using the χ^2-test. Moreover, a correct implementation of the algorithm is presented.

Hypothesis testing.

Neyman-Pearson paradigm somewhat rigid in that it requires to either reject or not reject the null hypothesis. Instead, might be interested in measuring strength of evidence against H_0.

Hypothesis testing.

Example: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ i.i.d. Test

$$H_0 \colon \mu = 0 \qquad \text{vs.} \qquad H_A \colon \mu = 1$$

with decision rule:

$$\text{reject } H_0 \text{ if } \bar{X}_n > c(\alpha) \text{, yielding level } \alpha \text{ test.}$$

On observing $\bar{X}_n = x$ we define

$$p^\star := \mathbb{P}\left\{ \bar{X}_n > x | H_0 \right\}.$$

Decision rule can now be equivalently written as

$$\begin{cases} \text{reject } H_0 & \text{if } p^\star \leq \alpha \\ \text{do not reject } H_0 & \text{otherwise} \end{cases}$$

that is $p^\star$ contains all the information we need about the sample to make the decision.

Hypothesis testing.

Definition (p-value)

The p-value is the smallest value of α for which the null hypothesis is rejected.

Remark

Suppose $T = T(X_1, \ldots, X_n)$ is continuous test statistic and rejection region is of the form

$$\{T > t\} \qquad \text{(for some \textit{critical value} t).}$$

Can interpret p-value

$$\mathbb{P}\{T(X_1, \ldots, X_n) > T(x_1, \ldots, x_n) | H_0\}$$

as probability under the null hypothesis of observing a value of the test statistic as or more extreme than the observation $T(x_1, \ldots, x_n)$.[4]

[4] Likewise for rejection region $\{T < t\}$.

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit). For the p-value we find

$$\mathbb{P}\left\{X^2 \geq 7.34\right\} = 0.025,$$

that is under the null hypothesis we would observe a value greater or equal to 7.34 in about 3 of 100 cases. In other words, the data suggest that H_0 is not very likely.