PH 712 Probability and Statistical Inference

Part IX: Hypothesis Testing

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2024/12/14 21:18:51

Is it a squirrel?

Figure 1: Flying Squirrel (Photograph by Joel Sartore)

- Make a decision between two hypotheses H_0 : YES and H_1: NO.
  - Checking necessary conditions under H_0
- It is a binary classification problem.

Problem formalization

- Assumptions
  - $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta)$
    - θ is fixed and unknown BUT is believed to be inside Θ
  - To make a decision on θ between two hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$
    - $\Theta_0 \cup \Theta_1 = \Theta$
    - $\Theta_0 \cap \Theta_1 = \emptyset$
- Four possible outcomes
  - True positive (TP): H_0 is wrong (i.e., H_1 is true) and we reject H_0 (i.e., accept H_1);
  - False positive (FP, type I error): H_0 is true (i.e., H_1 is wrong) but we reject H_0 (i.e., accept H_1);
  - True negative (TN): H_0 is true (i.e., H_1 is wrong) and we accept H_0 (i.e., reject H_1);
  - False negative (FN, type II error): H_0 is wrong (i.e., H_1 is true) but we accept H_0 (i.e., reject H_1).
  - E.g., in the context of identifying the animal,
    - TP: it is NOT a squirrel and is NOT identified as a squirrel
    - FP: it is a squirrel but is NOT identified as a squirrel
    - TN: it is a squirrel and is identified as a squirrel
    - FN: it is NOT a squirrel but is identified as a squirrel

	Accept H_0	Reject H_0
H_0 is true	True negative (TN)	False positive (FP, type I error)
H_0 is false	False negative (FN, type II error)	True positive (TP)

- Different objectives leading to different strategies:
  - Minimizing the misclassification rate: $\Pr(\text{FP}) + \Pr(\text{FN})$
    - Commonly adopted by classification techniques
  - Controlling the false discovery rate (FDR): $\Pr(\text{FP})/\{\Pr(\text{FP}) + \Pr(\text{TP})\}$
    - For sequential or simultaneous testing
  - Minimizing $\Pr(\text{FN})$ with $\Pr(\text{FP})$ capped; specifically, minimizing $\Pr(\text{type II error})$ with $\Pr(\text{type I error}) \leq \alpha$
    - Leading to the optimal hypothesis test

Formalizing the hypothesis test

- A test, say ϕ, is an indicator function

$$\phi(x_1, \ldots, x_n) = \mathbf{1}_R(x_1, \ldots, x_n) = \begin{cases} 0, & (x_1, \ldots, x_n) \notin R \\ 1, & (x_1, \ldots, x_n) \in R \end{cases}$$

  - Input: the sample or its realization
  - Output: the action after observing the input, i.e., 0 (accepting H_0) or 1 (rejecting H_0)
  - *Rejection region*: R, the set corresponding to the rejection of H_0
    - R is typically specified in terms of the realization of a *test statistic*; e.g., if $R = \{(x_1, \ldots, x_n) : \bar{x} \geq 3\}$, then $\bar{X}$ is a test statistic.
- Each test corresponds to a unique rejection region
  - Two tests are equivalent $\Leftrightarrow$ their rejection regions are identical

Uniformly most powerful (UMP) level α test (CB Sec 8.3.2)

- *Power function*: given a test ϕ and its rejection region R, the power function $\beta_\phi(\theta)$ is the probability of rejecting H_0: for all $\theta \in \Theta$,

$$\beta_\phi(\theta) = \Pr\{(X_1, \ldots, X_n) \in R\} = \Pr\{\phi(X_1, \ldots, X_n) = 1\}$$

  - $\Pr(\text{type I error}) = \beta_\phi(\theta)$ if θ is true AND $\theta \in \Theta_0$
  - $\Pr(\text{type II error}) = 1 - \beta_\phi(\theta)$ if θ is true AND $\theta \in \Theta_1$
  - Since the true θ is unknown, a good test requires small $\beta_\phi(\theta)$ for all $\theta \in \Theta_0$ AND large $\beta_\phi(\theta)$ for all $\theta \in \Theta_1$
- A test ϕ is of *size* $\alpha \Leftrightarrow \sup_{\theta \in \Theta_0} \beta_\phi(\theta) = \alpha$
  - $\sup_{\theta \in \Theta_0} \beta_\phi(\theta)$: the supremum of $\beta_\phi(\theta)$ in $\Theta_0 \Leftrightarrow$ the maximum of $\beta_\phi(\theta)$ in the closure of Θ_0
  - $\sup_{\theta \in \Theta_0} \beta_\phi(\theta) = \alpha \Rightarrow \Pr(\text{type I error}) \leq \alpha$
- A test ϕ is of *level* $\alpha \Leftrightarrow \sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha \Leftrightarrow$ the maximum of $\beta_\phi(\theta)$ in the closure of Θ_0
  - $\sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha \Rightarrow \Pr(\text{type I error}) \leq \alpha$
- Let ϕ be a level α test for $H_0 : \theta_0 \in \Theta_0$ vs $H_1 : \theta_0 \in \Theta_1$. If $\beta_\phi(\theta) \geq \beta_{\phi'}(\theta)$ for all $\theta \in \Theta_1$ and any other test ϕ' of level α, then ϕ is a UMP level α test.

Example Lec9.1

- $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$ with unknown θ and known σ. Consider a test for $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ with rejection region $\{(x_1, \ldots, x_n) : \sqrt{n}|\bar{x} - \theta_0|/\sigma > c\}$.
  1. Elaborate the power function.
  2. Find sample size n and threshold c if one desires that the type I error rate is 5% and the type II error rate at $\theta_0 + \sigma$ is 25%.

```r
N = 100
type2.err.rates = numeric(N)
c = -qnorm(.025)
f = function(n){
  pnorm(c-n^.5)-pnorm(-c-n^.5)
}
for (n in 1:N) {
  type2.err.rates[n] = f(n)
}
type2.err.rates # all the type II rates
min((1:N)[type2.err.rates<=.25]) # the min n such that the type II rate is lower than 25%
```

Likelihood ratio test (LRT, CB Sec. 8.2.1 & 10.3.1)

- Hypotheses: $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$
  - $\Theta = \Theta_0 \cup \Theta_1$
  - $\Theta_0 \cap \Theta_1 = \emptyset$
- Test statistic

$$\lambda(X_1, \dots, X_n) = \frac{L(\hat{\theta}_{\text{ML},0})}{L(\hat{\theta}_{\text{ML}})}$$

  - $\hat{\theta}_{\text{ML},0}$: MLE of θ under H_0
  - $\hat{\theta}_{\text{ML}}$: MLE of $\theta \in \Theta$
- Rejection region

$$R = \{(x_1, \dots, x_n) : \lambda(x_1, \dots, x_n) \le c_\alpha\},$$

  where c_α is chosen to make sure the size is α, i.e.,

$$\sup_{\theta \in \Theta_0} \beta_\phi(\theta) = \sup_{\theta \in \Theta_0} \Pr\{\lambda(X_1, \dots, X_n) \le c_\alpha\} = \alpha.$$

  - Essential but challenging to know the distribution of $\lambda(X_1, \dots, X_n)$ under H_0
- Implementation
  1. Confirm the value of α;
  2. Figure out $\hat{\theta}_{\text{ML},0}$ and $\hat{\theta}_{\text{ML}}$.
  3. Solve the following equation for c_α

$$\sup_{\theta \in \Theta_0} \beta_\phi(\theta) = \sup_{\theta \in \Theta_0} \Pr\{\lambda(X_1, \dots, X_n) \le c_\alpha\} = \alpha;$$

  4. Construct the rejection region $\{(x_1, \dots, x_n) : \lambda(x_1, \dots, x_n) \le c_\alpha\}$.
- Why is LRT promoted?
  - Neyman-Pearson Lemma (CB Thm 8.3.12): LRT is the UMP level α test for simple hypotheses ($H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$)
  - Karlin-Rubin theorem (CB Thm 8.3.17): under certain conditions, LRT is the UMP level α test for one-sided hypotheses ($H_0 : \theta \le \theta_0$ (or $\theta = \theta_0$) vs $H_1 : \theta > \theta_0$ OR $H_0 : \theta \ge \theta_0$ (or $\theta = \theta_0$) vs $H_1 : \theta < \theta_0$)
  - There is No UMP test for two-sided hypotheses ($H_0 : \theta = \theta_0$ vs $H_1 : \theta \ne \theta_0$) but LRT is UMP unbiased test for this scenario.
- Special cases
  - Equivalent to the Z-test if 1) the sample is iid normal with known variance and 2) the mean is to be tested
  - Equivalent to the t-test if 1) the sample is iid normal with unknown variance and 2) the mean is to be tested
  - Equivalent to the F-test if 1) the sample is iid normal with the mean and variance both unknown and 2) the variance is to be tested

LRT (con'd)

- Asymptotic rejection region (CB Thm 10.3.3)

  $$R \approx \{(x_1, \dots, x_n) : -2\ln\lambda(x_1, \dots, x_n) \geq \chi^2_{\nu,1-\alpha}\} = \{(x_1, \dots, x_n) : \lambda(x_1, \dots, x_n) \leq \exp(-\chi^2_{\nu,1-\alpha}/2)\},$$

  where $\chi^2_{\nu,1-\alpha}$ is the $(1-\alpha)$ quantile of $\chi^2(\nu)$, i.e., $F_{\chi^2(\nu)}(\chi^2_{\nu,1-\alpha}) = 1-\alpha$.
  - (CB Thm 10.3.1) Because, as $n \to \infty$, under H_0,

    $$-2\ln\lambda(X_1, \dots, X_n) \approx \chi^2(\nu),$$

    where $\nu =$ the difference of numbers of free parameters between Θ_0 and Θ.
- Implementation (asymptotic)
  1. Confirm the value of α;
  2. Figure out $\hat{\theta}_{\mathrm{ML},0}$ and $\hat{\theta}_{\mathrm{ML}}$;
  3. Check ν, the difference of numbers of free parameters between Θ_0 and Θ;
  4. Construct the asymptotic rejection region $\{(x_1, \dots, x_n) : -2\ln\lambda(x_1, \dots, x_n) \geq \chi^2_{\nu,1-\alpha}\}$.

Example Lec9.2

- Collecting sample $X_1, \dots, X_{1000} \overset{\text{iid}}{\sim} f_X(x \mid p) = p^x(1-p)^{1-x}$, $x = 0, 1$, $0 < p < 1$, we test $H_0 : p = 1/4$ vs. $H_1 : p \neq 1/4$.
  1. Derive the expression of test statistic $-2\ln\lambda(X_1, \dots, X_{1000})$, where $\lambda(X_1, \dots, X_{1000})$ is the likelihood ratio.
  2. Generate a figure (similar to CB Figure 10.3.1) to compare the simulated distribution of $-2\ln\lambda(X_1, \dots, X_{1000})$ and the chi-square approximation.

```
options(digits = 4)
set.seed(1)
B = 1e4L # time of replication
test_stats = numeric(B)
n = 1e3L # sample size
p0 = 1/4
for (i in 1:B){
  Xs = rbinom(n, 1, p0)
  Xbar = mean(Xs)
  test_stats[i] = 2*(n*Xbar*log(Xbar)+(n-n*Xbar)*log(1-Xbar)+n*Xbar*log(4)+(n-n*Xbar)*log(4/3))
}
seg = seq(0, 10, length.out=100)
pdfchi2 = dchisq(seg, 1)
hist(test_stats, breaks=100, xlim=c(0,10),
     freq=F, xlab = expression(paste('-2ln', lambda, '(x)')), main = '')
lines(seg, pdfchi2, col = "red")
```

CB Ex 8.2

- For a given city in a given year, assume that the number of automobile accidents follows a Poisson distribution. In past years the average number of accidents per year was 15, and this year it was 10. Is it justified to claim that the accident rate has dropped?

- Demo report: Testing hypotheses H_0 : ___ vs. H_1 : ___ , we carried out the ___ test and obtained ___ as the value of test statistic ___ . Since the rejection region is ___ , there was/wasn't a strong statistical evidence against H_0 at the ___ (significance) level, i.e., we believed that ___ .

- Adapting the demo report to this question: Testing hypotheses H_0 : accident rate $= 15$ vs. H_1 : accident rate < 15 , we carried out the likelihood ratio test and obtained 1.891 as the value of test statistic $30 + 2X(\ln X - \ln 15 - 1)$. Since the rejection region is $\{x : 30 + 2x(\ln x - \ln 15 - 1) \geq 3.841\}$, there wasn't a statistical evidence against H_0 at the .05 (significance) level, i.e., we believed that the accident rate hasn't dropped.

p-value (CB Sec 8.3.4)

- Motivation
  - Recall that a rejection region R consists of a test statistic (e.g., $\lambda(X_1, \ldots, X_n)$ for LRT) and critical point (e.g., c_α for LRT)
    * The test statistic NOT uniquely defined
    * The critical point varying with the definition of test statistic
  - Would like to fix the critical point to be α by defining a test statistic $p(X_1, \ldots, X_n)$ (i.e., p-value) such that the following set is equivalent to R

$$\{(x_1, \ldots, x_n) : p(x_1, \ldots, x_n) \leq \alpha\}$$

    * More convenient in communication because the critical point is α by default
- NOT always well-defined
- (CB Thm 8.3.27) If H_0 is rejected when the realization of test statistic $T(X_1, \ldots, X_n)$ is too large, then

$$p(x_1, \ldots, x_n) = \sup_{\theta \in \Theta_0} \Pr\{T(X_1, \ldots, X_n) \geq T(x_1, \ldots, x_n)\}.$$

  - $T(x_1, \ldots, x_n)$: the realization of test statistic $T(X_1, \ldots, X_n)$
- For LRT, asymptotically,

$$p(x_1, \ldots, x_n) = 1 - F_{\chi^2(\nu)}(-2\lambda(x_1, \ldots, x_n)).$$

  - $F_{\chi^2(\nu)}(\cdot)$: the cdf of $\chi^2(\nu)$

Revisit CB Ex 8.2

- For a given city in a given year, assume that the number of automobile accidents follows a Poisson distribution. In past years the average number of accidents per year was 15, and this year it was 10. Is it justified to claim that the accident rate has dropped?
- Demo report: Testing hypotheses H_0 : ___ vs. H_1 : ___ , we carried out the ___ test and obtained ___ as the p-value. So, at the ___ (significance) level, there was/wasn't a strong statistical evidence against H_0, i.e., we believed that ___ .

Wald test (CB pp. 493)

- Testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
- Test statistic: $(\hat{\theta}_{\text{ML}} - \theta_0)/\sqrt{\widehat{\text{var}}(\hat{\theta}_{\text{ML}})}$
  - Asymptotically equivalent to LRT for this pair of hypotheses
  - Refer to the previous part for how to obtain $\widehat{\text{var}}(\hat{\theta}_{\text{ML}})$ (via the Fisher information or delta methods)
- Level α Wald rejection region: $\{(x_1, \ldots, x_n) : |\hat{\theta}_{\text{ML}} - \theta_0|/\sqrt{\widehat{\text{var}}(\hat{\theta}_{\text{ML}})} \geq \Phi^{-1}_{1-\alpha/2}\}$
  - $\Phi^{-1}_{1-\alpha/2}$: the $(1 - \alpha/2)$ quantile of $\mathcal{N}(0, 1)$
- p-value $= 2\Phi\left(-|\hat{\theta}_{\text{ML}} - \theta_0|/\sqrt{\widehat{\text{var}}(\hat{\theta}_{\text{ML}})}\right)$
  - $\Phi(\cdot)$: cdf of $\mathcal{N}(0, 1)$

Score test (CB pp. 494)

- Testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$

- Test statistic: $\ell'(\theta_0)/\sqrt{I_n(\theta_0)}$ ($\approx \mathcal{N}(0,1)$ under H_0 as $n \to \infty$)
  - No need of MLE
- Level α score rejection region: $\{(x_1, \ldots, x_n) : |\ell'(\theta_0)|/\sqrt{I_n(\theta_0)} \geq \Phi^{-1}_{1-\alpha/2}\}$.
- p-value $= 2\Phi\left(-|\ell'(\theta_0)|/\sqrt{I_n(\theta_0)}\right)$

Revisit CB Ex 8.2

- For a given city in a given year, assume that the number of automobile accidents follows a Poisson distribution. In past years the average number of accidents per year was 15, and this year it was 10. Is it justified to claim that the accident rate has be changed?
- Demo report: Testing hypotheses H_0 : ___ vs. H_1 : ___ , we carried out the ___ test and obtained ___ as the p-value. So, at the ___ (significance) level, there was/wasn't a strong statistical evidence against H_0, i.e., we believed that ___ .