STAT 4100 Lecture Note

Week Twelve (Nov 28, 30 & Dec 2, 2022)

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca, zhiyanggeezhou.github.io)

2022/Dec/02 21:28:36

Approximation to the variance of $\hat{\theta}_n$

- Why?
  - Reflect the variation or dispersion of $\hat{\theta}_n$
  - Help approximate the distribution of $\hat{\theta}_n$ (and further construct the confidence region for θ) if assuming normality
- How?
  - Utilizing the asymptotic variance of $\hat{\theta}_n$
  - Resampling methods, e.g., bootstraping

CB Example 10.1.17 & Ex. 10.9 (con'd)

- iid $X_1, \ldots, X_n \sim p(x \mid \lambda) = \lambda^x \exp(-\lambda)/x!$, $x \in \mathbb{Z}^+$, $\lambda > 0$. Define $\theta = \Pr(X_i = 2 \mid \lambda) = \lambda^2 \exp(-\lambda)/2$. Approximate the variance of $\hat{\theta}_{\mathrm{ML}} = \bar{X}_n^2 \exp(-\bar{X}_n)/2$ by delta methods.

CB Example 10.1.15

- Holding iid $X_i \sim \mathrm{Bernoulli}(p)$, the variance of Bernoulli(p) is $\tau(p) = p(1-p)$ whose MLE is $\tau(\hat{p}_{\mathrm{ML}}) = \bar{X}_n(1-\bar{X}_n)$. Approximate $\mathrm{var}\{\tau(\hat{p}_{\mathrm{ML}})\}$ by delta methods.

Bootstraping the variance of $\hat{\theta}_n$ (CB Sec. 10.1.4)

- Nonparametric bootstrap:
  1. For j in $1:B$, do steps 2–3.
  2. Draw the jth resample $\boldsymbol{x}_j^*$ of size n from the original sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, with replacement, i.e., create a new iid sample $\boldsymbol{x}_j^*$ from F_n (the empirical cdf of the original sample)
  3. Let $\hat{\theta}_j^* = \hat{\theta}(\boldsymbol{x}_j^*)$.
  4. $\mathrm{var}(\hat{\theta}) \approx$ the sample variance of $\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$.
- (Optional, see, e.g., www.stat.columbia.edu/~bodhi/Talks/Emp-Proc-Lecture-Notes.pdf) Empirical process: theoretical foundation for nonparametric bootstrap
  - (Glivenko-Cantelli) $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0$
  - (Donsker) $\sqrt{n}(F_n - F) \xrightarrow{d} \mathrm{BB} \circ F$, i.e., $\mathrm{E}[g\{\sqrt{n}(F_n - F)\}] \to \mathrm{E}[g(\mathrm{BB} \circ F)]$ for all bounded, continuous and real-valued g
    - BB is a Gaussian process (specifically, standard Brownian bridge process on $[0,1]$), i.e.,
      - $\mathrm{BB}(0) = \mathrm{BB}(1) = 0$ but $\mathrm{BB}(t) \sim \mathcal{N}(0, t(1-t))$ for $t \in (0,1)$;

  - fixing $t_1, \dots, t_p \in (0,1)$, $[\mathrm{BB}(t_1), \dots, \mathrm{BB}(t_p)]^\top$ is of multivariate normal with $\mathrm{cov}(\mathrm{BB}(s), \mathrm{BB}(t)) = \min(s,t) - st$;
  - $\mathrm{BB}(t)$ is continuous in t.

```
options(digits = 4)
set.seed(1)
ts = (0:1000)/1000
delta_t = 1/1000
for (i in seq_len(animation::ani.options("nmax"))) {
  dev.hold()
  W = cumsum(c(0, rnorm(n = length(ts)-1, mean = 0, sd = delta_t^.5)))
  BB = W-ts*W[length(W)]
  plot(y = BB, x = ts, xlim = c(0,1), ylim = c(-1.5,1.5), type='l', xlab = 't', ylab = 'BB(t)',
       main = paste('Sample path', i, 'of the standard Brownian Bridge on [0,1]'))
  abline(h = 0, lty = 2)
  animation::ani.pause()
}
```

- Parametric bootstrap:
  1. For j in $1:B$, do steps 2–3.
  2. Draw the jth resample $\boldsymbol{x}_j^*$ of size n from a fitted model $f(x \mid \hat\theta)$.
  3. Let $\hat\theta_j^* = \hat\theta(\boldsymbol{x}_j^*)$.
  4. $\mathrm{var}(\hat\theta) \approx$ the sample variance of $\{\hat\theta_1^*, \dots, \hat\theta_B^*\}$.

CB Example 10.1.15

- Holding iid $X_i \sim \mathrm{Bernoulli}(p)$, the variance of Bernoulli(p) is $\tau(p) = p(1-p)$ for which the MLE is $\tau(\hat p_{\mathrm{ML}}) = \bar X_n(1-\bar X_n)$. Approximate $\mathrm{var}\{\tau(\hat p_{\mathrm{ML}})\}$ by the bootstrap.

```
options(digits = 4)
set.seed(1)
B = 1e4L
n = 30
x = rbinom(n, 1, prob = .7)
theta_ml = mean(x)
tau_theta_star_np = numeric(B)
tau_theta_star_p = numeric(B)
# Nonparametric bootstrap
for (j in 1:B){
  x_star = sample(x, size = n, replace = T)
  tau_theta_star_np[j] = mean(x_star)*(1-mean(x_star))
}
var(tau_theta_star_np)
# Parametric bootstrap
for (j in 1:B){
  x_star = rbinom(n, size = 1, prob = theta_ml)
  tau_theta_star_p[j] = mean(x_star)*(1-mean(x_star))
}
var(tau_theta_star_p)
# Estimate via the first-order delta method
theta_ml*(1-theta_ml)*(1-2*theta_ml)^2/n
# Estimate via the second-order delta method
2*theta_ml^2*(1-theta_ml)^2/n^2
```

Large-sample hypothesis testing

Recall the LRT

- $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ v.s. $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$, where $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1$
- LRT statistic
$$\lambda(\boldsymbol{x}) = \frac{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} L(\boldsymbol{\theta}; \boldsymbol{x})}{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}; \boldsymbol{x})} = \frac{L(\hat{\boldsymbol{\theta}}_{0,\mathrm{ML}}; \boldsymbol{x})}{L(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}; \boldsymbol{x})}$$
  - $\hat{\boldsymbol{\theta}}_{0,\mathrm{ML}}$: constrained MLE for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$
  - $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$: unconstrained MLE for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
- $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \le c_\alpha\}$: rejection region of level α LRT
  - c_α is such defined that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} \Pr(\lambda(\mathbf{X}) \le c_\alpha \mid \boldsymbol{\theta}) = \alpha$

Asymptotic LRT rejection region (CB Thm 10.3.1 & 10.3.3)

- Under H_0, as $n \to \infty$,
$$-2\ln\lambda(\mathbf{X}) \xrightarrow{d} \chi^2(\nu),$$
  where ν = difference of numbers of free parameters in $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}$.
- (CB Thm 10.3.3) $\{\boldsymbol{x} : -2\ln\lambda(\boldsymbol{x}) \ge \chi^2_{\nu,1-\alpha}\}$: asymptotic rejection region of level α LRT
  - $\chi^2_{\nu,1-\alpha}$ is the $1-\alpha$ quantile of $\chi^2(\nu)$.

CB Example 10.3.4

- iid $X_1, \dots, X_n \sim f(x \mid p_1, \dots, p_5) = p_x$, $x = 1, \dots, 5$, $\sum_{k=1}^5 p_k = 1$ and $p_k \in (0, 1)$. i.e., the categorical distribution. Specify the asymptotic level α LRT rejection region for $H_0 : p_1 = p_2 = p_3$ and $p_4 = p_5$ vs. H_1 : Otherwise.

```
options(digits = 4)
set.seed(1)
B = 1e4L
n = 1e3L
p0 = c(.1, .1, .1, .35, .35)
ys = rmultinom(B, n, p0)
ms = matrix(
  c(rep(colMeans(ys[1:3,]), times = 3), rep(colMeans(ys[4:5,]), times = 2)),
  nrow = length(p0), ncol = B, byrow = T)
test_stats = 2*colSums(ys*log(ys/ms))
seg = seq(0, 20, length.out=1000)
pdfchi2 = dchisq(seg, 3)
hist(test_stats, breaks=100, xlim=c(0,20),
     freq=F, xlab = expression(paste('-2ln', lambda, '(x)')), main = '')
lines(seg, pdfchi2, col = "red")
```

Take-home exercises (NOT to be submitted; to be potentially covered in labs)

- CB Ex. 10.17(a–c), 10.36, 10.38
- HMC Ex. 6.3.16–6.3.18