STA 114: Statistics

Lab 2

This lab work is intended to illustrate the foundation of the classical theory of statistics based on sampling theory. But before we go there, we need to learn how to approximate the probability of a certain event through computer simulation. The technique used below is known as the Monte Carlo approximation, and it essentially relies of the strong laws of large numbers we learn in basic probability courses.

To begin with, if you flip an unloaded coin an awfully large number of times, then the number of times you'd see a head is roughly 50%. Weak and strong laws of large numbers say that this percentage would indeed converge to the 50% mark as you keep on flipping.

Evaluating probability and expectation through computer simulation

More generally, if you have a random variable X with pdf/pmf $f(x)$ and you want to calculate $P(X \in A)$ or $\mathrm{E}g(X)$ for some set A or some function g, then you could simulate a lot of random numbers $x_1, x_2, \cdots, x_M$ from the pdf/pmf $f(x)$ and approximate:

$$P(X \in A) \approx \frac{1}{M}\#\{x_i \in A, i = 1, \cdots, M\}, \;\; \mathrm{E}g(X) \approx \frac{1}{M}\sum_{i=1}^{M} g(x_i),$$

and approximations get better for larger M. With modern computers, it is fairly straightforward to simulate random numbers from a pdf/pmf, and so the above strategy is a very practically useful one for evaluating probabilities of expectations.

To test this, consider $X \sim \mathsf{Normal}(0,1)$ and suppose we want to computer $P(X > 1)$ and $\mathrm{E}e^{X/3}$. From direct integrations these two quantities are $P(X > 1) = 1 - \Phi(1) = 0.1587$ and $\mathrm{E}e^{X/3} = e^{1/18} = 1.0571$ (here $\Phi(x)$ denotes the standard normal CDF). Below are Monte Carlo approximations to these quantities.

```
> M <- 10000
> x <- rnorm(M)
> mean(x > 1)
[1] 0.1556
> mean(exp(x/3))
[1] 1.055558
```

Clearly, the approximation works well. The error in approximation is typically of the order $1/\sqrt{M}$.

> Task 1. Approximate $P(|X| > 1.96)$ and $\mathrm{E}X^2$ for $X \sim \mathsf{Normal}(0,1)$. Are the approximation errors of the order $1/\sqrt{M}$?

Coverage probability

For $X \sim \mathsf{Binomial}(n, p)$, suppose we want to evaluate the coverage probability of

$$B_c(x) = \frac{x}{n} \mp \frac{c}{\sqrt{n}} \sqrt{\frac{x}{n}\left(1 - \frac{x}{n}\right)}$$

at a fixed point of interest p_0. The following code gives an approximation:

```
coverage <- function(n, p0, c, M = 10000){
     x <- rbinom(M, n, p0)
     Bc.left <- x / n - c / sqrt(n) * sqrt((x/n) * (1 - x/n))
     Bc.rght <- x / n + c / sqrt(n) * sqrt((x/n) * (1 - x/n))
     include.p0 <- (Bc.left < p0) & (Bc.rght > p0)
     return(mean(include.p0))
}
```

To run this code, copy and paste it on your R console. And then issue the command:

```
> coverage(500, 0.5, 1.96)
[1] 0.9473
```

TASK 2. Use this code to evaluate $\gamma(B_c; p_0)$, with $= n = 500$ and $p_0 = 0.5$ for $c = 1/2, 1, 1.96, 2, 3, 4, 5$. Then repeat this with $p_0 = 0.25$ and $p_0 = 0.9$. Make a comparison between the three sets as to how close $\gamma(B_{1.96}; p_0)$ is to 0.95.

Square error risk of mean and median: Normal model

Consider $X_1, X_2, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, 1)$ where $\mu \in (-\infty, \infty)$ is an unknown parameters. We shall compare the two estimators $\bar{X}$ and X_{med} of μ through square error risks $r(\bar{X}; \mu_0) = \mathrm{E}_{[X|\mu_0]}(\bar{X} - \mu_0)^2$ and $r(X_{\text{med}}; \mu_0) = \mathrm{E}_{[X|\mu_0]}(X_{\text{med}} - \mu_0)^2$ at $\mu_0 = 1.37$, with $n = 50$. A Monte Carlo version is:

```
> M <- 10000
> n <- 50
> mu0 <- 1.3
> x.mean <- replicate(M, mean(rnorm(n, mu0, 1)))
> risk.Xbar <- mean((x.mean - mu0)^2)
```

TASK 3. Run this code and record the value of `risk.Xbar`. Compare it with the exact value $1/n$.

TASK 4. Modify the code and run it for X_{med} and record `risk.Xmed`. Compare it with the "large sample (i.e., n) approximation" $\pi/(2n)$.

TASK 5. Record which one is a better estimator of μ at $\mu = 1.3$.

Square error risk of mean and median: Laplace model

Now consider the case where $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Laplace}(x|\mu, 1)$, $\mu \in (-\infty, \infty)$ and we want to compare $\bar{X}$ and X_{med} as estimators of μ.

> TASK 6. Repeat tasks 3-5 for this model.

The exact value of $r(\bar{X}; \mu_0) = 2/n$ and a large sample approximation of $r(X_{\text{med}}; \mu_0)$ is σ^2/n. Use the following code to simulate from the Laplace distribution.

```
rlap <- function(n, mu = 0, sigma = 1){
  u <- runif(n)
  return(mu + sigma * sign(u - .5) * log(2 * pmin(u, 1 - u)))
}
```

Interval based on the median

Again consider $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, 1)$, $\mu \in (-\infty, \infty)$. We know that when $\mu = \mu_0$, the median X_{med} has mean μ_0 and variance approximately $\pi/(2n)$. Consider the interval $A(x) = x_{\text{med}} \mp 1.96\sqrt{\pi/(2n)}$ for μ.

> TASK 7. Evaluate the coverage $\gamma(A; 1.3)$ with $n = 50$.

Begin your code as follows

```
M <- 10000
n <- 50
mu0 <- 1.3
x.med <- replicate(M, median(rnorm(n, mu0, 1)))
```