14.381: Statistics

Problem Set 1

due Tuesday, September 17, 2019, at lecture

You should hand in the solution for problem 3 to our TA(David). Problems 1-2 and 4-6 are for practice. They will be discussed at recitation.

1. Let X and Y be random variables with finite variances.

   (i) Show that
   $$\min_{g(\cdot)} E\left(Y - g(X)\right)^2 = E\left(Y - E\left(Y \mid X\right)\right)^2,$$
   where $g(\cdot)$ ranges over all functions.

   (ii) Assume $m(X) = E(Y|X)$ and write $Y = m(X) + e$. Show that $Var(Y) = Var(m(X)) + Var(e)$.

   (iii) If $E(Y|X = x) = a + bx$ find $E(YX)$ as a function of moments of X.

2. Show that if a sequence of random variables ξ_i converges in distribution to a constant c, then $\xi_i \xrightarrow{p} c$.

3. *(The required problem)* Let $\{X_i\}$ be independent Bernoulli (p). Then $EX_i = p$, $Var(X_i) = p(1-p)$. Let $Y_n = \frac{1}{n}\sum_{i=1}^n X_i$.

   (a) Describe the asymptotic behavior of Y_n.

   (b) Show that for $p \neq \frac{1}{2}$ the estimated variance $Y_n(1-Y_n)$ has the following limit behavior
   $$\sqrt{n}(Y_n(1-Y_n) - p(1-p)) \Rightarrow N(0, (1-2p)^2 p(1-p)).$$

   (c) Prove that if (i) $\frac{\sqrt{n}}{\sigma}(\xi_n - \mu) \Rightarrow N(0,1)$ (ii) g is twice continuously differentiable: $g'(\mu) = 0$, $g''(\mu) \neq 0$, then
   $$n(g(\xi_n) - g(\mu)) \Rightarrow \sigma^2 \frac{g''(\mu)}{2}\chi_1^2.$$

Note. You may assume that g has more derivatives, if it simplifies your life. Use O_p and o_p notation wherever possible.

Note: χ_1^2 is a chi-square distribution with 1 degree of freedom. Let $\xi_1, \ldots, \xi_p$ be i.i.d. $N(0,1)$, then $\chi_p^2 = \sum_{i=1}^p \xi_i^2$.

(d) Show that for $p = \frac{1}{2}$

$$n\left[Y_n(1-Y_n) - \frac{1}{4}\right] \Rightarrow -\frac{1}{4}\chi_1^2$$

Curious fact: Note that $Y_n(1-Y_n) \leq \frac{1}{4}$, that is, we always underestimate the variance for $p = \frac{1}{2}$.

4. Prove the following statements:

   (a) If $X_n = O_p(n^{-\delta})$ for some $\delta > 0$ then $X_n = o_p(1)$;

   (b) If $X_n = o_p(b_n)$ then $X_n = O_p(b_n)$;

   (c) If $X_n = O_p(n^\alpha)$ and $Y_n = o_p(n^\beta)$, then $X_n Y_n = o_p(n^{\alpha+\beta})$.

5. (Computer experiment) Let $X_1, ..., X_n$ be i.i.d. $N(0,1)$ random variables and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Plot $\bar{X}_n$ versus n for $n = 1, ..., 10000$. Repeat for $X_1, X_2, ..X_n$ distributed i.i.d. Cauchy ($f(x) = \frac{1}{\pi(1+x^2)}$ for $x \in \mathbb{R}$). Explain why there is such a difference.

6. (Computer experiment) This experiment is intended to support your derivations in problem 3. First, fix $n = 100$ and perform the following simulation

   (i) Simulate $X_i, i = 1, ..., n$ from a Bernoulli distribution with parameter p.

   (ii) Calculate the statistic $Z = \widehat{p}(1-\widehat{p})$, where $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i$.

   (iii) Store the value of Z, then repeat steps (i) and (ii) for $B = 1000$ times.

   (iv) Draw the histogram of Z, along with the expected asymptotic distribution.

   Repeat the experiment above for $p = 0.5, 0.48$, and 0.4. Repeat each of the three cases for $n = 1000$. Comment on how well the approximation works in each case, and try to explain the simulation results.