18.650. Statistics for Applications
Fall 2016. Problem Set 1

Due Friday, Sep. 16 at 12 noon

Problem 1 Convergence of random variables

1. For $n \in \mathbb{N}^*$, let X_n be a random variable such that $\mathbb{P}\left[X_n = \frac{1}{n}\right] = 1 - \frac{1}{n^2}$ and $\mathbb{P}[X_n = n] = \frac{1}{n^2}$. Does X_n converge in probability? In L^2? $P\{|X_n - 0| > \varepsilon\} \to 0$
2. Let $(X_n)_{n\in\mathbb{N}^*}$ be a sequence of i.i.d. Bernoulli random variables with parameter $p \in [0, 1]$. For $n \in \mathbb{N}^*$, let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
   a) What is the distribution of $n\bar{X}_n$? $C_n^k p^k (1-p)^{n-k}$
   b) Prove that $\bar{X}_n$ converges to p in L^2?
3. For $n \in \mathbb{N}^*$, let X_n be a Poisson random variable with parameter $1/n$.
   a) Prove that $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$. $P\{X_n = k\} = \frac{(\frac{1}{n})^k}{k!} e^{-\frac{1}{n}}$
   b) Prove that $nX_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$.

Problem 2 True or false

For each of the following statement, say whether it is true or false. When your answer is "false", give a counter example.

1. If $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$ and $Y_n \xrightarrow[n\to\infty]{\text{a.s.}} Y$, then $X_n + Y_n \xrightarrow[n\to\infty]{\text{a.s.}} X + Y$. T
2. If $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ and $Y_n \xrightarrow[n\to\infty]{\mathbb{P}} Y$, then $X_n + Y_n \xrightarrow[n\to\infty]{\mathbb{P}} X + Y$. T
3. If $X_n \xrightarrow[n\to\infty]{(d)} X$ and $Y_n \xrightarrow[n\to\infty]{(d)} Y$, then $X_n + Y_n \xrightarrow[n\to\infty]{(d)} X + Y$. F ex: $Y = -X$
4. Consider a coin that shows Heads with some unknown probability p when it is tossed. After tossing this coin 100 times, Heads have shown up 43 times. The unknown parameter p is contained in the interval $[.33, .53]$ with probability 95%.

Problem 3 A confidence interval for Bernoulli random variables

Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli random variables with unknown parameter $p \in (0, 1)$ and denote by $\bar{X}_n$ their empirical average:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

1. Show that $\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}}$ converges in distribution to a standard Gaussian random variable Z. $\sim \mathcal{N}(0,1)$
2. Prove that for all $t > 0$,

$$\mathbb{P}[|Z| \leq t] = 2\mathbb{P}[Z \leq t] - 1.$$

3. For $t > 0$, let $\mathcal{I}_t$ be the interval

$$I_t = \left[\bar{X}_n - \frac{t\sqrt{p(1-p)}}{\sqrt{n}}, \bar{X}_n + \frac{t\sqrt{p(1-p)}}{\sqrt{n}}\right].$$

Using the previous questions, prove that

$$\mathbb{P}[I_t \ni p] \to 2\Phi(t) - 1, \quad n \to \infty,$$

where Φ is the cumulative distribution function of the standard Gaussian distribution.

4. In practice, we would like to be able to define an interval as small as possible, whose expression does not depend on the unknown value of p. Using the previous question, find the value of t such that the interval I_t contains p with probability going to 95% as n grows to infinity. Denote by t_0 this value.
Hint: The 97.5%*-quantile of the standard Gaussian distribution is* 1.96.

In the next questions, we study three ways of updating the interval I_{t_0} in order to obtain intervals that no longer depend on the unknown value of p.

5. A conservative method:
   a) Prove that no matter the value of p,

$$p(1-p) \leq \frac{1}{4}.$$

$\left[\bar{X}_n - \frac{t}{2\sqrt{n}}, \bar{X}_n + \frac{t}{2\sqrt{n}}\right]$

   b) Using the previous question, find an interval J_1 centered around $\bar{X}_n$ that does not depend on p but that contains p with probability at least 95% when n becomes large.
   c) If it was known a priori that $p \leq .3$, could you find an interval smaller than J_1, independent of p and that still contains p with probability at least 95% when n becomes large ?

$\left[\bar{X}_n \pm \frac{t\sqrt{0.3 \cdot 0.7}}{\sqrt{n}}\right]$

6. Solving an inequality:
   a) Prove that the statement "$p \in I_{t_0}$" is equivalent to a polynomial inequality of degree 2 in p.

b) Solve this inequality in p.

c) Using the previous question, propose a new interval J_2 that does not depend on p but contains p with probability going to 95% as n goes to infinity.

7. Show that if you replace p with $\bar{X}_n$ in the expression of I_{t_0}, you obtain an interval J_3 that contains p with probability going to 95% as n goes to infinity.

$$\sqrt{n}\frac{\bar X_n - p}{\sqrt{p(1-p)}} = \sqrt{n}\frac{\bar X_n - p}{\sqrt{\bar X_n(1-\bar X_n)}} \cdot \left(\frac{\sqrt{\bar X_n(1-\bar X_n)}}{\sqrt{p(1-p)}}\right)$$

$$\downarrow$$

$$1$$

8. In order to forecast the results of the elections, n registered voters are sampled randomly and with replacement and are asked whether they plan to vote for Hillary or Donald at the presidential elections. Denote by p the proportion of American voters who plan to vote for Hillary, within the whole population and by $\hat{p}$ the proportion of sampled individuals who answer that they plan to vote for Hillary.

   a) Assume that $n = 10,000$ and $\hat{p} = .7341$. Compute the intervals J_1, J_2 and J_3 defined in the previous questions and compare their respective lengths.

   b) Find a minimal sample size n that would be required in order to be able to compute an interval of length at most .05 that does not depend on the unknown value of p and that contains p with approximate probability at least 95%.