S.-T. Yau College Student Mathematics Contests 2019

Probability and Statistics

Individual (4 problems)

1) Suppose $(X_n)_{n\geq 1}$ is a sequence of positive random variables. There exists a constant $C > 0$ such that,

$$\mathbb{E}[X_n] \leq C, \quad \mathbb{E}[\max\{0, -\log X_n\}] \leq C, \quad \forall n.$$

Show that

$$\limsup_{n\to\infty} X_n^{1/n} = 1.$$

2) Suppose γ is a probability measure on $\{0, 1, 2\}$ such that $\gamma(0) > \gamma(2) > 0$. Let $(\xi_n)_{n\geq 1}$ be a sequence of i.i.d. random variables with common law γ. Define the sequence

$$Y_0 = 0, \quad Y_{n+1} = \max\{0, Y_n + \xi_{n+1} - 1\}, \quad \forall n \geq 0.$$

Show that $(Y_n)_{n\geq 0}$ is an irreducible Markov chain on the state space $\mathbb{N} = \{0, 1, 2, \ldots\}$ and it is positive recurrent.

3) Suppose $(\epsilon_n)_{n\geq 1}$ is a sequence of i.i.d. random variables and the common law is Bernoulli:

$$\mathbb{P}[\epsilon_1 = 1] = \mathbb{P}[\epsilon_1 = -1] = 1/2.$$

Consider the random series $f(x) = \sum_{n=1}^{\infty} \epsilon_n x^n$. Show that the random series attains zero infinitely many times on $x \in [0, 1)$ almost surely.

4) Consider a randomized experiment with $2N$ units, half to be randomly assigned an active treatment, and the other half to be assigned the control treatment; the objective is to measure the effect of the active versus control treatments on an outcome, called Y. For example, the units could be people with high blood pressure, where Y is blood pressure one week after receiving the active drug or an inactive drug, a placebo, where the patient is blinded to which drug is being given.

The estimand, the goal of the experiment, is the average value of Y if all $2N$ units received active minus the average value of Y if all $2N$ units received control. Assume that these $2 \times 2N$ numbers are fixed quantities (in the statistical literature this assumption is known as SUTVA the stable-unit-treatment-value assumption), which means that the outcome Y for the i-th unit receiving a particular treatment is a proper function of that unit and the treatment that unit i received.

Derive the following results in this simple situation.

a) Find the expectation of the estimator, the difference in the observed sample means (between those assigned treatment and those assigned control), in terms of the estimand (defined above), where expectation in this context refers to averaging over all possible random allocations.

b) The variance of the estimator described in part a) (again, with variance defined as averaging over all possible random allocations).

c) Find an unbiased estimator of the variance in part b), assuming additive treatment effects, that is, the treatment minus control values of Y are constant across the $2N$ units, so that the treatment versus control condition adds a constant value for all $2N$ units.

d) Find the bias of the estimator in part c) when the treatment effects are non-additive.

e) Generalize the results in parts a),b), c) and d) to the situation where $2N$ is replaced by $N_t + N_c$, with N_t units getting active treatment and N_c units getting control, where these sample sizes are unequal.

f) Argue that the estimator in part c), when the sample sizes are large, will look gaussian, and conduct a small simulation to indicate that this often happens with relatively small sample sizes.

g) Modify the first four parts to consider a different randomized experiment, but still with $2N$ units, half to be allocated to active and half to be allocated to control, but now we have a covariate, X, a background variable that is suspected to be related to Y. For example, X could be blood pressure today, pre-treatment. In this experiment, many randomized allocations are considered, but all allocations are rejected if the sample X means of the treated and controls are too different for example more than a standard deviation apart. Be careful here to note which results generalize and which do not.

h) Finally, consider the randomized experiment in part g) when N_t units are treated and Nc are control, where N_t and N_c are not equal: Which results in parts a),b),c) and d) generalize without modification? In particular, describe how the conclusion in part f) changes. Note that this is an interesting situation where the asymptotic distributions of sample means are not gaussian. What are these distributions?