STATS 300A Theory of Statistics
Stanford University, Fall 2015

Problem Set 1

Due: Thursday, October 1, 2015 **Reading:** TSH 1.1-1.2, 1.4; K 3.1-3.3

Instructions:

- You may appeal to any result proved in class or proved in the course textbooks.
- Any request to find a statistic with certain properties requires proof that those properties are satisfied.

Definition 1. A statistic A is *ancillary* for $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$ if the distribution of $A(X)$ does not depend on the model parameter θ.

Definition 2. A statistic T is *complete* for $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$ if $\mathbb{E}_\theta [f(T(X))] = 0, \forall \theta \in \Omega \Rightarrow f(T(X)) = 0$ a.s. under each $\mathbb{P}_\theta \in \mathcal{P}$.

Problem 1. (Sufficiency) You're interested in which factors contribute to the popularity of a YouTube video, so you write a script to scrape $p = 10,000$ features (e.g., time since posting, use of autotuning, presence of cats) from each video on YouTube, along with the the number of views at the time of scraping. You call the feature vector associated with the i-th video $x_i \in \mathbb{R}^p$ and the associated number of views $y_i \in \mathbb{R}$. For simplicity, you decide to model each (x_i, y_i) pair as an independent observation with x_i drawn from a known distribution q_i and the conditional distribution of $y_i \mid x_i$ given by $\mathcal{N}(\langle x_i, \beta \rangle, 1)$ where $\beta \in \mathbb{R}^p$ is unknown. You quickly realize that you cannot afford to store the data for all $n \geq 3$ billion videos on your laptop.

(a) Use the Neyman-Fisher factorization criterion to show that a statistic of dimension $p^2 + p$ (or smaller) is sufficient for making inferences about β.

(b) Describe how you would compute your statistic from part (a) in an online (and hence memory-efficient) fashion, that is, assuming that you are only given access to a single datapoint (x_i, y_i) at a time and that (x_i, y_i) must be discarded before the next datapoint is viewed.

Problem 2. (Exponential Families) Suppose that X belongs to an s-dimensional exponential family with densities of the form

$$p(x; \eta) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x),$$

for h and each T_i continuously differentiable and η in the natural parameter space Θ.

(a) Prove the following variant of *Stein's identity*: Suppose that

(i) the support of X is $(-\infty, \infty)$,

(ii) g is continuously differentiable with $\mathbb{E}\left[|g(X)|\right] < \infty$, $\mathbb{E}\left[|g'(X)|\right] < \infty$, and $\mathbb{E}\left[\left|\left(\frac{h'(X)}{h(X)} + \sum_{i=1}^{s} \eta_i T_i'(X)\right) g(X)\right|\right] < \infty$.

Then

$$\mathbb{E}\left[\left(\frac{h'(X)}{h(X)} + \sum_{i=1}^{s} \eta_i T_i'(X)\right) g(X)\right] = -\mathbb{E}\left[g'(X)\right].$$

Stein's identity (which holds under even weaker assumptions) is the basis of many modern model selection procedures, distributional convergence arguments, and concentration inequalities.

(b) Under the conditions of part (a), show that $\mathbb{E}\left[g(X)(X-\mu)\right] = \sigma^2 \mathbb{E}\left[g'(X)\right]$ whenever $X \sim N(\mu, \sigma^2)$.

(c) Compute the third and fourth (uncentered) moments of the $N(\mu, \sigma^2)$ distribution using part (b).

Problem 3. (Minimal Sufficiency) As founder of a new social networking service, you observe the presence or absence of a friendship link between each pair of your users. You decide to model these observations as independent Bernoulli draws $X_{ij} \sim \text{Bernoulli}(p_{ij})$, where $X_{ij} = 1$ indicates the presence of a friendship link between distinct users $i < j \in \{1, \ldots, n\}$, and $X_{ij} = 0$ indicates an absence.

(a) Find a minimal sufficient statistic for the model parameterization $\theta = (p_{ij})_{i<j}$.

(b) Experience tells you that each user has a propensity for interaction with others that largely determines social network interactions. To capture this intuition, you consider a constrained, lower-dimensional model family, in which

$$p_{ij} = \frac{\exp\left(\beta_i + \beta_j\right)}{1 + \exp\left(\beta_i + \beta_j\right)}$$

for link-propensity parameters $\beta_i \in \mathbb{R}$. Find a minimal sufficient statistic for the model parameterization $\theta = (\beta_1, \ldots, \beta_n)$.

(c) Prove the sufficiency (just sufficiency, not minimal sufficiency) of your statistic from part (b) again, this time using the definition of sufficiency directly.

Problem 4. (Ancillarity) Consider an i.i.d. sample $X_1, \ldots, X_n \in \mathbb{R}$ from a *location-scale family* specified by the cumulative distribution function $F_{a,b}(x) \triangleq F((x-a)/b)$. Here, F is a known c.d.f., while the real numbers a and b are the location and scale parameters of the family respectively.

(a) If b is known, show that the differences $(X_1 - X_i)/b$ for $i = 2, \ldots, n$ are ancillary.

(b) If a is known, show that the differences $(X_1 - a)/(X_i - a)$ for $i = 2, \ldots, n$ are ancillary.

(c) If neither a nor b is known, show that the ratios $(X_1 - X_i)/(X_2 - X_i)$ for $i = 3, \ldots, n$ are ancillary.

Problem 5. (Medley) Consider an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the two-dimensional normal distribution

$$N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \theta\\ \theta & 1\end{bmatrix}\right),$$

with $\theta \in \Omega = (-1, 1)$.

(a) Find a two-dimensional minimal sufficient statistic.

(b) Prove that the minimal sufficient statistic in (a) is not complete.

(c) Prove that $Z_1 = \sum_{i=1}^n X_i^2$ and $Z_2 = \sum_{i=1}^n Y_i^2$ are both ancillary, while (Z_1, Z_2) is not ancillary.

Problem 6. How much time did you spend on this problem set?