STOR 654 Homework 11

1. Show carefully that the total variation distance $\mathrm{TV}(P,Q)$ is a metric.

2. Show carefully that the Hellinger distance $\mathrm{H}(P,Q)$ is a metric.

3. Suppose that we observe $Y \sim \mathcal{N}_n(\theta, I)$ and define one-sided p-values $p_i = \overline{\Phi}(Y_i)$ for $1 \leq i \leq n$. Consider the two-step testing procedure

- Reject the global null H_0 if $\min p_i \leq \alpha/n$
- If we reject H_0 then reject $\mathrm{H}_{0,i}$ if $p_i \leq \alpha$.

(a) Show that this procedure gives weak control of the FWER.

(b) Show that if $\theta_j \geq (1+\epsilon)\sqrt{2\log n}$ for one or more indices j then $V(p) \approx \alpha \, |\{i : \theta_i = 0\}|$

(c) Argue informally that the procedure does not provide strong control of the FWER.

4. Let $\gamma_1, \gamma_2 : [0,1]^n \to \{0,1\}^n$ be multiple testing procedures. The procedure γ_1 is said to be more conservative than the procedure γ_2 if for each $p \in [0,1]^n$ we have $\gamma_1(p) \leq \gamma_2(p)$ where the inequality is componentwise. Equivalently, the set of hypotheses rejected by γ_1 is always a subset of the those rejected by γ_2.

(a) Show that the Bonferroni multiple testing procedure is more conservative than Holm's step-down procedure.

(b) Is Holm's step-down procedure more conservative than the Benjamini-Hochberg step-up procedure? Justify your answer.

5. Let $X_1, \ldots, X_n$ be non-negative random variables such that $\sum_{i=1}^n X_i$ is equal to a non-negative constant T. Show that at least $2n/3$ of the expectations $\mathbb{E}X_1, \ldots, \mathbb{E}X_n$ are less than or equal to $3T/n$.

6. Let $X_1, \ldots, X_n$ be iid Rademacher (sign) variables with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$.

(a) Using the variance bound from an earlier HW, show that X_i has maximum variance among all random variables supported on $[-1,1]$.

(b) Identify the common moment generating function $M_X(s)$ of the X_i, which is a simple sum of exponentials.

(c) Establish the bound $M_X(s) \leq e^{s^2/2}$. Hint: Expand the exponentials in $M_X(s)$, cancel identical terms, and examine the coefficients of the remaining terms.

(d) Use the MGF bound in part (c) and Chernoff's probability bound to find an upper bound on $\mathbb{P}(\sum_{i=1}^n X_i \geq t)$ for $t \geq 0$.

(e) Use Hoeffding's inequality to bound the probability in part (d) and compare the bound you found there. Comment.

7. (Pinsker's inequality) Pinsker's inequality relates the L_1 distance between two density function to their Kullback-Liebler divergence. It has many uses in statistics and probability. Here we derive Pinsker's inequality from a numerical inequality and Cauchy-Schwartz.

a. Show that for $x \geq 0$ one has the inequality

$$(x-1)^2 \ \leq \ \left(\frac{4+2x}{3}\right)(x \log x - x + 1)$$

Hint: Let $g(x)$ be the difference between the right- and left-hand sides of the inequality. Expand $g(x)$ in a third order Taylor series around $x = 1$.

b. Let f and g be probability density functions. Establish Pinsker's inequality

$$\int |f(x) - g(x)| \, dx \ \leq \ \sqrt{2\mathrm{KL}(f : g)}$$

Hint: Note that the right hand side can be written as $\int |f/g - 1| \, g \, dx$. Apply the square root form of the inequality above to the integrand and then apply Cauchy-Schwarz.

8. Symmetric formulas for the variance

(a) Let X be a random variable with finite second moment, and let X' be an independent copy of X (a random variable that is independent of X but has the same distribution). Show that $\mathrm{Var}(X) = \frac{1}{2}\mathbb{E}(X - X')^2$.

(b) Let $x = x_1, \ldots, x_n \in \mathbb{R}$ be a finite sample with average $\overline{x}$. Show that

$$\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2 \ = \ \frac{1}{2n(n-1)}\sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2$$

(c) Use the formula in (b) to show that the sample variance $\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is unbiased.

9. (Log-Sum Inequality)

(a) Show that the function $f(x) = x \log x$ is strictly convex.

(b) Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be positive numbers. Show that

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality iff all the ratios a_i/b_i are equal. Hint: Define a discrete random variable X with values $x_i = a_i/b_i$ and probability mass function $p(x_i) = b_i/\sum_j b_j$. Apply Jensen's inequality to the expectation of $X \log X$.