Statistics 581, Problem Set 6 Solutions

Wellner; 11/9/2017

1. Suppose that $X_1, \ldots, X_n$ are i.i.d. with distribution function F which has positive density f at its quartiles $F^{-1}(1/4)$ and $F^{-1}(3/4)$ and at its median $F^{-1}(1/2)$.
(a) Let $M_n = \mathbb{F}_n^{-1}(1/2)$ be the median and $Q_n = \mathbb{F}_n^{-1}(3/4) - \mathbb{F}_n^{-1}(1/4)$, the interquartile range. Find the joint asymptotic distribution of M_n and Q_n as estimators of the population median $m = M(F) = F^{-1}(1/2)$ and interquartile-quartile range $q = Q(F) = F^{-1}(3/4) - F^{-1}(1/4)$. That is, prove that
$$\sqrt{n}(M_n - m, Q_n - q)^T \to_d \text{“something”}$$
and find “something”.
(b) Assuming that the underlying distribution F is Cauchy(μ, σ) ($X = \sigma Y + \mu$ where $Y \sim$ Cauchy$(0,1)$), express μ and σ in terms of $m = M(F)$ and $q = Q(F)$.
(c) Use the relations you derived in (b) to propose estimators of μ and σ based on M_n and Q_n. Show that your estimators $\hat{\mu}_n$ and $\hat{\sigma}_n$ satisfy
$$\sqrt{n}(\hat{\mu}_n - \mu, \hat{\sigma}_n - \sigma) \to_d N_2(0, \Sigma)$$
and compute Σ.
Hint: A standard Cauchy random variable has d.f. $G(x) = 1/2 + (1/\pi)\arctan(x)$ and inverse distribution function $G^{-1}(u) = \tan(\pi(u - 1/2))$.

Solution: (a) First note that by Theorem 7.2, chapter 2, course notes,
$$\underline{Y}_n \equiv \sqrt{n} \begin{pmatrix} \mathbb{F}_n^{-1}(1/2) - F^{-1}(1/2) \\ \mathbb{F}_n^{-1}(1/4) - F^{-1}(1/4) \\ \mathbb{F}_n^{-1}(3/4) - F^{-1}(3/4) \end{pmatrix} \to_d \begin{pmatrix} Q'(1/2)\mathbb{V}(1/2) \\ Q'(1/4)\mathbb{V}(1/4) \\ Q'(3/4)\mathbb{V}(3/4) \end{pmatrix} \equiv \underline{Y} \sim N_3(0, \Sigma)$$
where $\mathbb{V} = -\mathbb{U}$ is a standard Brownian bridge process on $[0,1]$ and where, with $(Q'(1/2), Q'(1/4), Q'(3/4)) \equiv (a, b, c)$,
$$\Sigma = \text{diag}(a,b,c)\frac{1}{16}\begin{pmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{pmatrix}\text{diag}(a,b,c).$$
Now
$$\sqrt{n}(M_n - m) = \sqrt{n}(\mathbb{F}_n^{-1}(1/2) - F^{-1}(1/2)) = Y_{n,1},$$
and
$$\sqrt{n}(Q_n - q) = \sqrt{n}(\mathbb{F}_n^{-1}(3/4) - \mathbb{F}_n^{-1}(1/4) - (F^{-1}(3/4) - F^{-1}(1/4))) = Y_{n,3} - Y_{n,2}.$$
Thus
$$\begin{aligned} \sqrt{n}\begin{pmatrix} M_n - m \\ Q_n - q \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix}\underline{Y}_n \to_d \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix}\underline{Y} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix}\text{diag}(a,b,c)\begin{pmatrix} \mathbb{V}(1/2) \\ \mathbb{V}(1/4) \\ \mathbb{V}(3/4) \end{pmatrix} \\ &= \begin{pmatrix} a\mathbb{V}(1/2) \\ c\mathbb{V}(3/4) - b\mathbb{V}(1/4) \end{pmatrix} \\ &\sim N_2(0, \widetilde{\Sigma}) \end{aligned}$$

where

$$\widetilde{\Sigma} \;=\; \frac{1}{16}\begin{pmatrix} 4a^2 & 2a(c-b) \\ 2a(c-b) & 3c^2-2bc+3b^2 \end{pmatrix}.$$

(b) When F is the Cauchy (μ,σ) distribution, $F(x) = G((x-\mu)/\sigma)$, $Q(u) = \mu + \sigma G^{-1}(u) = \mu + \sigma\tan(\pi(u-1/2))$. Thus

$$\begin{aligned} m &= m(F) = \mu + \sigma G^{-1}(1/2) = \mu + \sigma\cdot\tan(0) = \mu, \\ q &= Q(F) = F^{-1}(3/4) - F^{-1}(1/4) = \mu + \sigma G^{-1}(3/4) - (\mu + \sigma G^{-1}(1/4)) \\ &= \sigma(\tan(\pi/4) - \tan(-\pi/4)) = 2\sigma. \end{aligned}$$

(c) It follows from (b) together with (a) that we can take $\hat{\mu}_n = M_n = \mathbb{F}_n^{-1}(1/2)$ and $\hat{\sigma}_n = 2^{-1}Q_n = 2^{-1}(\mathbb{F}_n^{-1}(3/4) - \mathbb{F}_n^{-1}(1/4))$. In this case $Q'(u) = \sigma\pi/\cos(\pi(u-1/2))^2$, and we compute

$$\begin{aligned} &a = Q'(1/2) = \sigma\pi, \\ &b = Q'(1/4) = \sigma\pi/\cos^2(-\pi/4) = 2\pi\sigma, \\ &c = Q'(3/4) = \sigma\pi/\cos^2(+\pi/4) = 2\pi\sigma, \\ &2a(c-b) = 0, \\ &3c^2 - 2bc + 3b^2 = 2\cdot 3(2\pi\sigma)^2 - 2(2\pi\sigma)^2 = 16\pi^2\sigma^2. \end{aligned}$$

It follows that

$$\begin{aligned} \sqrt{n}\begin{pmatrix} \hat{\mu}_n - \mu \\ \hat{\sigma}_n - \sigma \end{pmatrix} &= \sqrt{n}\begin{pmatrix} M_n - \mu \\ 2^{-1}(Q_n - q) \end{pmatrix} \\ &\to_d \frac{\pi\sigma}{4} N_2\left(0, \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right). \end{aligned}$$

2. (a) Let $\mathbb{U}_X$ and $\mathbb{U}_Y$ be two independent Brownian bridge processes on $[0,1]$, and let $\lambda \in [0,1]$. Show that the process $\mathbb{U}$ defined by $\mathbb{U} = \sqrt{1-\lambda}\mathbb{U}_X - \sqrt{\lambda}\mathbb{U}_Y$ is also a Brownian bridge process.

(b) Suppose that $X_1,\ldots,X_m$ are i.i.d. F and $Y_1,\ldots,Y_n$ are i.i.d. G, with the X's and Y's independent. Let $\mathbb{F}_m$ and $\mathbb{G}_n$ denote the empirical df's of the X's and Y's respectively. Suppose that $\lambda_N \equiv m/N \to \lambda \in (0,1)$ where $N \equiv m+n$. Show that

$$\begin{aligned} \mathbb{X}_{m,n} &\equiv \sqrt{\frac{mn}{N}}(\mathbb{F}_m - F) - \sqrt{\frac{mn}{N}}(\mathbb{G}_n - G) \\ &\Rightarrow \sqrt{1-\lambda}\mathbb{U}_X(F) - \sqrt{\lambda}\mathbb{U}_Y(G) \equiv \mathbb{X}. \end{aligned}$$

(c) A distribution function F is said to be *stochastically smaller than another distribution function* G, and we write $F <_s G$, if $F(x) \geq G(x)$ for all $x \in \mathbb{R}$ with strict inequality for some $x \in \mathbb{R}$. Note that this means $F^{-1}(u) \leq G^{-1}(u)$ for all $0 < u < 1$ so that the random variables resulting from a Skorokhod construction with one uniform random variable ξ satisfy satisfy $X^* \equiv F^{-1}(\xi) \leq G^{-1}(\xi) \equiv Y^*$. Consider testing $H_0:\ F = G$ continuous versus $H_1:\ F <_s G$ based on the one-sided Kolmogorov-Smirnov statistic

$$D^+_{m,n} \equiv \sqrt{\frac{mn}{N}}\|(\mathbb{F}_m - \mathbb{G}_n)^+\|_\infty = \sqrt{\frac{mn}{N}}\sup_{x\in\mathbb{R}}(\mathbb{F}_m(x) - \mathbb{G}_n(x));$$

here the notation f^+ is the *positive part* of the function f: $f^+(x) \equiv \max\{f(x), 0\}$. Use the result of B to show that under H_0 it follows that

$$D^+_{m,n} \to_d \|\mathbb{U}^+\|_\infty = \sup_{0\le t\le 1} \mathbb{U}(t).$$

(d) To test the effectiveness of vitamin B_1 in stimulating growth in mushrooms, vitamin B_1 was applied to 13 mushrooms selected at random from a group of 24, while the remaining 11 did not receive this treatment. The weights of the mushrooms at the end of the period of observation were:

$\underline{X} = (18, 14.5, 13.5, 12.5, 23, 24, 21, 17, 18.5, 9.5, 14)$, $m = 11$;
$\underline{Y} = (27, 34, 20.5, 29.5, 20, 28, 20, 26.5, 22, 24.5, 34, 35.5, 19)$, $\quad n = 13$.

Plot the two empirical df's and compute $D^+_{m,n}$. What is the approximate P - value for testing H_0 versus $H_1 : \ F <_s G$? You may use your favorite tables of the distribution of $D^+_{m,n}$, or the asymptotic distribution.

Solution: A. Let $\mathbb{Z} \equiv \sqrt{1-\lambda}\mathbb{U}_X - \sqrt{\lambda}\mathbb{U}_Y$. Then

$$\begin{aligned} E\mathbb{U}(t) &= \sqrt{1-\lambda}E\mathbb{U}_X(t) - \sqrt{\lambda}E\mathbb{U}_Y(t) = \sqrt{1-\lambda}\cdot 0 - \sqrt{\lambda}\cdot 0 = 0, \\ Cov[\mathbb{U}(s), \mathbb{U}(t)] &= (1-\lambda)Cov(\mathbb{U}_X(s), \mathbb{U}_X(t)) + \lambda Cov(\mathbb{U}_Y(s), \mathbb{U}_Y(t)) \\ &\qquad \text{since } \ \mathbb{U}_X, \mathbb{U}_Y \ \text{ are independent} \\ &= (1-\lambda)(s\wedge t - st) + \lambda(s\wedge t - st) = s\wedge t - st. \end{aligned}$$

Thus $\mathbb{U}$ is a mean zero Gaussian process with covariance $s \wedge t - st$; it follows that $\mathbb{U}$ is a Brownian bridge process on $[0, 1]$.
B. This follows from the continuous mapping (Mann - Wald) theorem: with $\mathbb{Z}_m \equiv \sqrt{m}(\mathbb{F}_m - F)$, $\mathbb{W}_n \equiv \sqrt{n}(\mathbb{G}_n - G)$, and $\lambda_N \equiv m/(m+n) \equiv m/N$, we have $(\mathbb{Z}_m, \mathbb{W}_n, \lambda_N) \Rightarrow (\mathbb{Z}, \mathbb{W}, \lambda)$ where $\mathbb{Z} \equiv \mathbb{U}_X(F)$ and $\mathbb{W} \equiv \mathbb{U}_Y(G)$ are independent F and G Brownian bridge processes on $\mathbb{R}$ respectively. Since the function $g : D(\mathbb{R}) \times D(\mathbb{R}) \times \mathbb{R} \mapsto D(\mathbb{R})$ defined by

$$g(z, w, \lambda) = \sqrt{1-\lambda}z - \sqrt{\lambda}w$$

is continuous (with respect to $\|\cdot\|_\infty \vee \|\cdot\|_\infty \vee |\cdot|$), it follows that $g(\mathbb{Z}_m, \mathbb{W}_n, \lambda_N) \Rightarrow g(\mathbb{Z}, \mathbb{W}, \lambda)$.

Here is an alternative argument (using Skorokhod constructions) that makes the continuity of g more explicit: For special constructions we have

$$\sqrt{m}(\mathbb{F}_m - F) \stackrel{d}{=} \mathbb{U}^*_{X,m}(F) \qquad \text{and} \qquad \sqrt{n}(\mathbb{G}_n - G) \stackrel{d}{=} \mathbb{U}^*_{Y,n}(G)$$

where the starred processes satisfy

$$\|\mathbb{U}^*_{X,m}(F) - \mathbb{U}^*_X(F)\|_\infty \to_{a.s.} 0, \qquad \text{and} \qquad \|\mathbb{U}^*_{Y,n}(G) - \mathbb{U}^*_X(G)\|_\infty \to_{a.s.} 0;$$

here $\mathbb{U}^*_X$ and $\mathbb{U}^*_Y$ are independent Brownian bridge processes Thus, with

$$\begin{aligned} \mathbb{X}^*_{m,n} &\equiv \sqrt{1-\lambda_N}\mathbb{U}^*_{X,m}(F) - \sqrt{\lambda_N}\mathbb{U}_{Y,n}(G), \\ \mathbb{X}^*_N &\equiv \sqrt{1-\lambda_N}\mathbb{U}^*_X(F) - \sqrt{\lambda_N}\mathbb{U}^*_Y(G), \text{and} \\ \mathbb{X}^* &\equiv \sqrt{1-\lambda}\mathbb{U}^*_X(F) - \sqrt{\lambda}\mathbb{U}^*_Y(G), \end{aligned}$$

it follows that

$$\begin{aligned}\|\mathbb{X}^*_{m,n} - \mathbb{X}^*_N\|_\infty &= \|\sqrt{1-\lambda_N}(\mathbb{U}^*_{X,m}(F) - \mathbb{U}^*_X(F)) - \sqrt{\lambda_N}(\mathbb{U}^*_{Y,n}(G) - \mathbb{U}^*_Y(G)\|_\infty \\ &\leq \sqrt{1-\lambda_N}\|\mathbb{U}^*_{X,m}(F) - \mathbb{U}^*_X(F)\|_\infty + \sqrt{\lambda_N}\|\mathbb{U}^*_{Y,n}(G) - \mathbb{U}^*_Y(G)\|_\infty \\ &\to_{a.s.} 0 + 0 = 0. \end{aligned} \quad (0.1)$$

Furthermore

$$\|\mathbb{X}^*_N - \mathbb{X}^*\|_\infty \leq |\sqrt{1-\lambda_N} - \sqrt{1-\lambda}|\|\mathbb{U}_X\|_\infty + |\sqrt{\lambda_N} - \sqrt{\lambda}|\mathbb{U}_Y\|_\infty \to_{a.s.} 0. \quad (0.2)$$

Combining (0.1) and (0.2) using the triangle inequality yields

$$\|\mathbb{X}^*_{m,n} - \mathbb{X}^*\|_\infty \to_{a.s.} 0.$$

C. Under $H_0 : \ F = G$ continuous, it follows that

$$\begin{aligned} D_{m,n} &= \|\sqrt{\frac{mn}{N}}(\mathbb{F}_m - \mathbb{G}_n)^+\|_\infty \\ &= \|\left\{\sqrt{\frac{mn}{N}}(\mathbb{F}_m - F) - \sqrt{\frac{mn}{N}}(\mathbb{G}_n - G)\right\}^+\|_\infty \\ &\to_{a.s.} \|(\mathbb{X}^*)^+\|_\infty = \|(\mathbb{Z}^*)^+\|_\infty \stackrel{d}{=} \|\mathbb{U}^+\|_\infty \end{aligned}$$

where $\mathbb{U}$ is a Brownian bridge process. Hence $D^+_{m,n} \to_d \|\mathbb{U}^+\|_\infty = \sup_{0<t<1}\mathbb{U}(t)$.
D. A quick look at a plot of the empirical distributions $\mathbb{F}_m$ and $\mathbb{G}_n$ convinces one that there is indeed a difference, so there is relatively little need for the formal test. Nevertheless, the observed significance level or $P-$value gives some indication of the magnitude of the difference. I compute $D^+_{m,n} = \sqrt{(11 \cdot 13)/24}(8/11) \approx 1.7752\ldots$ for the observed data; the supremum is achieved on the interval $[18.5, 19) = [X_{(8)}, Y_{(1)})$. This yields an approximate $P-$value of $\exp(-2(1.7752)^2) \approx .0018\ldots$.

Unfortunately, tables of the exact distribution of $D^+_{m,n}$ for small sample sizes $m \neq n$ do not seem to exist. The best available tables for $D_{m,n}$ for $m \neq n$ with $m \leq n \leq 100$ are those of Kim and Jennrich in *Selected Tables in Mathematical Statistics, Volume I*, pages 79 - 170, American Mathematical Society, Providence, 1970, H. L. Harter and D. B. Owen, editors. The original tables of Massey (1952), *Ann. Math. Statist.* **23**, for $D^+_{m,n}$ and $D_{m,n}$ only go through $m, m \leq 10$. It is also true that the convergence in distribution of the two-sample Kolmogorov-Smirnov statistic occurs quite quickly for $m = n \to \infty$, but slowly for $m \neq m \wedge n \to \infty$; this was studied carefully by J. L. Hodges (1958), "The significance probability of the Smirnov two-sample test", *Archiv for Mathematik* **3**, 469 - 486. What seems to be needed is a simple and reliable algorithm implementing the graphical recursion methods of Hodges for small $m \neq n$; see e.g. J. D. Gibbons (1971), *Nonparametric Statistical Inference*, pages 128 - 131.

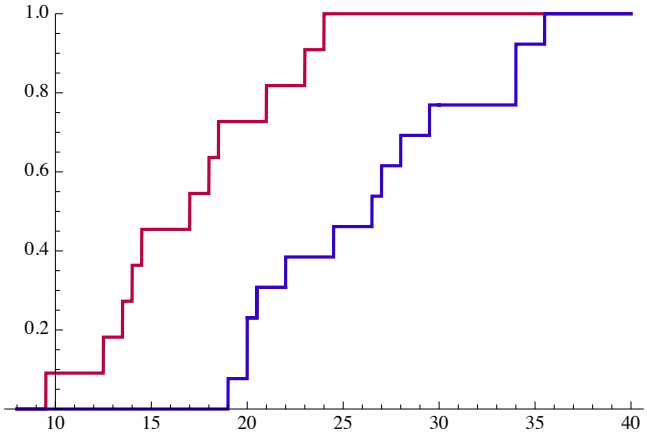

Figure 1: The empirical distribution $\mathbb{F}_m$ (of X's, red), and empirical distribution $\mathbb{G}_n$ of (Y's, blue)

3. A. Compute and plot the score for location $-f'/f(x)$ when:
(a) $f = \phi$, the standard normal density;
(b) $f(x) = \exp(-x)/(1+\exp(-x))^2$ (logistic);
(c) $f(x) = (1/2)\exp(-|x|)$ (double exponential);
(d) $f(x) = t_k$, the $t-$density with $k-$degrees of freedom;
(e) $f(x) = \exp(-x)\exp(-\exp(-x))$;
(f) $f(x) = 2\phi(x)\Phi(ax)$ where $\Phi(x)$ is the standard normal distribution function and $a > 0$: correction here! I missed a factor of 2 in the density in the problems statement.
B. A density f is called *log-concave* if $\log f$ is a concave function. Which of the densities in (a) - (e) are log-concave?

Solution: A. (a) For the normal density $f = \phi$, $I_f = \int x^2\phi(x)dx = Var(Z) = 1$ where $Z \sim N(0,1)$.
(b) For the logistic density the information for location is

$$\begin{aligned} I_f &= \int_{-\infty}^{\infty} \left(\frac{1-e^{-x}}{1+e^{-x}}\right)^2 dF(x) \\ &= \int_{-\infty}^{\infty} (2F(x)-1)^2 dF(x) = \int_0^1 (2u-1)du = 4Var(U) \\ &= 4/12 = 1/3 \end{aligned}$$

where $U \sim$ Uniform$(0,1)$.
(c) For the double exponential density $-(f'/f)^2(x) = 1$, so $I_f = 1$.
(d) For the t_k density, by using a change of variables and letting T_r denote a random variable with the t_r distribution with r degrees of freedom,

$$\begin{aligned} I_f &= \int_{-\infty}^{\infty} \left(\frac{k+1}{k}\right)^2 \frac{x^2}{(1+x^2/k)^2} \frac{\Gamma((k+1)/2)}{\Gamma(k/2)\sqrt{\pi k}} \frac{1}{(1+x^2/k)^{(k+1)/2}} dx \\ &= \frac{(k+1)(k+2)}{(k+4)(k+3)} Var(T_{k+4}) \\ &= \frac{(k+1)(k+2)}{(k+4)(k+3)} \frac{k+4}{k+2} \end{aligned}$$

$$= \frac{k+1}{k+3}$$

since $Var(T_r) = r/(r-2)$ for $r > 2$.
(e) For the extreme value distribution, $F(x) = \exp(-\exp(-x))$, and therefore if $X \sim F$ the random variable $Y = \exp(-X) \sim$ exponential(1):

$$P(Y \geq y) = P(\exp(-X) \geq y) = P(X \leq -\log(y)) = \exp(-\exp(\log y)) = \exp(-y).$$

Since $-(f'/f)(x) = -1 + e^{-x}$, it is easily seen that

$$I_f = E\{(-f'/f)^2(X)\} = E\{(e^{-X} - 1)^2\} = E(Y-1)^2 = Var(Y) = 1.$$

(f) For the skewed normal distribution $f(x) = 2\phi(x)\Phi(ax)$ with $a \neq 0$
(slight correction here: I missed the factor of 2 in stating the problem!)
$\log f(x) = \log \phi(x) = \log(2 \cdot (2\pi)^{-1/2}) - x^2/2 + \log \Phi(ax)$ and we find that

$$-\frac{f'}{f}(x) = x + \frac{a\phi(ax)}{\Phi(ax)}.$$

I do not know an explicit expression for the Fisher information for location.
B. The normal, logistic, double exponential, and Gumbel densities in (a), (b), (c), (e), and (f) are all log-concave with monotone increasing score functions for location. The t_k density in (d) is *not* log-concave. But the t_k density is $s-$concave with $s = -1/(k+1)$ since with $f_k(x) = c_k(1 + x^2/k)^{-(k+1)/2}$ we have

$$f_k(x)^s = c_k^2 \cdot (1 + x^2/k)^{1/2},$$

which is a convex function (of x).